

Asynchronous Adaptation and Learning over Networks — Part III: Comparison Analysis

Xiaochuan Zhao, *Student Member, IEEE*, and Ali H. Sayed, *Fellow, IEEE*

Abstract

In Part II [3] we carried out a detailed mean-square-error analysis of the performance of asynchronous adaptation and learning over networks under a fairly general model for asynchronous events including random topologies, random link failures, random data arrival times, and agents turning on and off randomly. In this Part III, we compare the performance of synchronous and asynchronous networks. We also compare the performance of decentralized adaptation against centralized stochastic-gradient (batch) solutions. Two interesting conclusions stand out. First, the results establish that the performance of adaptive networks is largely immune to the effect of asynchronous events: the mean and mean-square convergence rates and the asymptotic bias values are not degraded relative to synchronous or centralized implementations. Only the steady-state mean-square-deviation suffers a degradation in the order of ν , which represents the small step-size parameters used for adaptation. Second, the results show that the adaptive distributed network matches the performance of the centralized solution. These conclusions highlight another critical benefit of cooperation by networked agents: cooperation does not only enhance performance in comparison to stand-alone single-agent processing, but it also endows the network with remarkable resilience to various forms of random failure events and is able to deliver performance that is as powerful as batch solutions.

Index Terms

Distributed optimization, diffusion adaptation, asynchronous behavior, centralized solutions, batch solutions, adaptive networks, dynamic topology, link failures.

The authors are with Department of Electrical Engineering, University of California, Los Angeles, CA 90095 Email: xiaochuanzhao@ucla.edu, sayed@ee.ucla.edu.

This work was supported by NSF grants CCF-1011918 and ECCS-1407712. A short and limited early version of this work appeared in the conference proceeding [1]. The first two parts of this work are presented in [2], [3].

I. INTRODUCTION

In Part I [2] we introduced a general model for asynchronous behavior over adaptive networks that allowed for various sources of uncertainties including random topologies, random link failures, random data arrival times, and agents turning on and off randomly. We showed that despite these uncertainties, which could even occur simultaneously, the adaptation process remains mean-square stable for sufficiently small step-sizes. Specifically, we derived condition (93) in Part I [2], namely,

$$\frac{\bar{\mu}_k^{(2)}}{\bar{\mu}_k^{(1)}} < \frac{\lambda_{k,\min}}{\lambda_{k,\max}^2 + \alpha} \quad (1)$$

for all k , to ensure that the steady-state individual mean-square-deviations (MSD) satisfies

$$\limsup_{i \rightarrow \infty} \mathbb{E} \|w^o - \mathbf{w}_{k,i}\|^2 = O(\nu) \quad (2)$$

for all k , where $\bar{\mu}_k^{(m)} \triangleq \mathbb{E}[\boldsymbol{\mu}_k(i)]^m$ denotes the m -th moment of the random step-size parameter $\boldsymbol{\mu}_k(i)$, $\{\lambda_{k,\min}, \lambda_{k,\max}, \alpha\}$ are from Assumptions 2 and 3 of Part I [2], w^o denotes the optimal minimizer, and

$$\nu \triangleq \max_k \frac{\sqrt{\bar{\mu}_k^{(4)}}}{\bar{\mu}_k^{(1)}} \quad (3)$$

Note that in Theorem 1 from Part I [2], we used ν_o in (2), where ν_o is from (95) of Part I [2]. Since $\nu_o \leq \nu$ by (107) from Part I [2], we replaced ν_o with ν in (2).

In Part II [3] we examined the attainable mean-square-error (MSE) performance of the asynchronous network and derived expressions that reveal how close the estimates at the various agents get to the desired optimal solution that is sought by the network. In particular, we showed among other results that under a strengthened condition (19) from Part II [3] (relative to condition (1)), namely,

$$\frac{\sqrt{\bar{\mu}_k^{(4)}}}{\bar{\mu}_k^{(1)}} < \frac{\lambda_{k,\min}}{3\lambda_{k,\max}^2 + 4\alpha} \quad (4)$$

for all k , it holds that

$$\limsup_{i \rightarrow \infty} \mathbb{E} \|\mathbf{w}_{k,i} - \mathbf{w}_{\ell,i}\|^2 = O(\nu^{1+\gamma'_o}) \quad (5)$$

for all k and ℓ , where $\gamma'_o > 0$ is some constant that is given by (92) of Part II [3].

In (200) and (201) from Appendix E of Part I [2], we showed that condition (4) implies condition (1) so that both results (2) and (5) hold. Expressions (2) and (5) show that all agents are able to reach a level of $O(\nu^{1+\gamma'_o})$ agreement with each other and to get $O(\nu)$ close to w^o in steady-state. These results establish that asynchronous networks can operate in a stable manner under fairly general asynchronous events and, importantly, are able to adapt and learn well. Two important questions remain to be addressed:

- 1) Compared with synchronous networks, does the asynchronous behavior degrade performance?
- 2) How close can the performance of an asynchronous network get to that of a centralized solution?

In this Part III, we therefore compare the performance of synchronous and asynchronous networks. We also compare the performance of distributed solutions against centralized (batch) solutions. The results will show that the performance of adaptive networks are surprisingly immune to the effect of asynchronous uncertainties: the mean and mean-square convergence rates and the asymptotic bias values are not degraded relative to synchronous or centralized implementations. Only the steady-state MSD suffers a degradation of the order of ν . The results also show that an adaptive network always matches the performance of a centralized solution. The main results of this part are summarized in Table I, which compares various performance metrics across different implementations. The notation in Table I will be explained in the sequel. For now, we simply remark that the results in Table I show that the distributed and centralized implementations have almost the same mean-square performance in either the synchronous or asynchronous modes of operation, i.e., the asynchronous distributed implementation approaches the asynchronous centralized implementation, and the synchronous distributed implementation approaches the synchronous centralized implementation.

We indicated in the introductory remarks of Part I [2] that studies exist in the literature that examine the performance of distributed strategies in the presence of some forms of asynchronous uncertainties [4]–[7] or changing topologies [5]–[13], albeit for decaying step-sizes. We also explained how the general asynchronous model that we introduced in Part I [2] covers broader situations of practical interest, including adaptation and learning under constant step-sizes, and how it allows for the simultaneous occurrence of multiple random events from various sources. Still, these earlier studies do not address the two questions posed earlier on how asynchronous networks compare in performance to synchronous networks and to centralized (batch) solutions. If it can be argued that asynchronous networks are still able to deliver performance similar to synchronous implementations where no uncertainty occurs, or similar to batch solutions where all information is aggregated and available for processing in a centralized fashion, then such a conclusion would be of significant practical relevance. The same conclusion would provide a clear theoretical justification for another critical benefit of cooperation by networked agents, namely, that cooperation does not only enhance performance in comparison to stand-alone single-agent processing, as already demonstrated in prior works in the literature (see, e.g., [14]–[16] and the references therein), but it also endows the network with remarkable resilience to various forms of uncertainties and is still able to deliver performance that is as powerful as batch solutions.

TABLE I
COMPARISON OF SYNCHRONOUS VS. ASYNCHRONOUS AND DISTRIBUTED VS. CENTRALIZED SOLUTIONS

	Synchronous Distributed	Asynchronous Distributed	Synchronous Centralized	Asynchronous Centralized
Algs.	(88a) and (88b)	(71a) and (71b)	(66)	(7)
Vars. ^a	$\mathbf{w}_{i,\text{sync}}^{\text{diff}}$	$\mathbf{w}_{i,\text{async}}^{\text{diff}}$	$\mathbf{w}_{i,\text{sync}}^{\text{cent}}$	$\mathbf{w}_{i,\text{async}}^{\text{cent}}$
Paras. ^b	$\{\bar{a}_{\ell k}, \bar{\mu}_k\}$	$\{\mathbf{a}_{\ell k}(i), \boldsymbol{\mu}_k(i)\}$	$\{\bar{\pi}_k, \bar{\mu}_k\}$	$\{\boldsymbol{\pi}_k(i), \boldsymbol{\mu}_k(i)\}$
Mn. Rate ^c	$\rho(\bar{\mathcal{B}}) = \rho_o + O(\nu^{1+1/N})$	$\rho(\bar{\mathcal{B}}) = \rho_o + O(\nu^{1+1/N})$	$\rho(\bar{B}) = \rho_o$	$\rho(\bar{B}) = \rho_o$
M.S. Rate ^d	$\rho(\mathcal{F}_{\text{sync}}) = \rho_o^2 + O(\nu^{1+1/N})$	$\rho(\mathcal{F}_{\text{async}}) = \rho_o^2 + O(\nu^{1+1/N^2})$	$\rho(F_{\text{sync}}) = \rho_o^2$	$\rho(F_{\text{async}}) = \rho_o^2 + O(\nu^2)$
MSD ^e	$\frac{1}{4}\text{Tr}(H^{-1}R_{\text{sync}}) + O(\nu^{1+\gamma_o})$	$\frac{1}{4}\text{Tr}(H^{-1}R_{\text{async}}) + O(\nu^{1+\gamma_o})$	$\frac{1}{4}\text{Tr}(H^{-1}R_{\text{sync}}) + O(\nu^{1+\gamma_o})$	$\frac{1}{4}\text{Tr}(H^{-1}R_{\text{async}}) + O(\nu^{1+\gamma_o})$

^a Variables. The variables for synchronous diffusion strategies are denoted in the table by $\mathbf{w}_{i,\text{sync}}^{\text{diff}} \triangleq \text{col}\{\mathbf{w}_{1,i,\text{sync}}^{\text{diff}}, \mathbf{w}_{2,i,\text{sync}}^{\text{diff}}, \dots, \mathbf{w}_{N,i,\text{sync}}^{\text{diff}}\}$, where $\mathbf{w}_{k,i,\text{sync}}^{\text{diff}}$ denotes the iterate of agent k at time i . The variables for asynchronous diffusion strategies are defined in the same manner.

^b Parameters. The parameters used by the four strategies satisfy:

- 1) $\bar{\mu}_k = \mathbb{E}[\boldsymbol{\mu}_k(i)]$.
- 2) $c_{\mu,k,\ell} = \mathbb{E}[(\boldsymbol{\mu}_k(i) - \bar{\mu}_k)(\boldsymbol{\mu}_\ell(i) - \bar{\mu}_\ell)]$.
- 3) $\bar{a}_{\ell k} = \mathbb{E}[\mathbf{a}_{\ell k}(i)]$, $\bar{\pi}_k = \mathbb{E}[\boldsymbol{\pi}_k(i)]$, and $\bar{\pi}_k = \bar{p}_k$, where $\bar{A} = [\bar{a}_{\ell k}]_{\ell,k=1}^N$, $\bar{p} = [\bar{p}_k]_{k=1}^N$, $\bar{A}\bar{p} = \bar{p}$, and $\bar{p}^\top \mathbf{1}_N = 1$.
- 4) $c_{a,\ell k,nm} = \mathbb{E}[(\mathbf{a}_{\ell k}(i) - \bar{a}_{\ell k})(\mathbf{a}_{nm}(i) - \bar{a}_{nm})]$, $c_{\pi,k,\ell} = \mathbb{E}[(\boldsymbol{\pi}_k(i) - \bar{\pi}_k)(\boldsymbol{\pi}_\ell(i) - \bar{\pi}_\ell)]$, and $C_\pi = P_p - \bar{p}\bar{p}^\top$, where $C_A = [c_{a,\ell k,nm}]_{\ell,k,n,m=1}^N$, $C_\pi = [c_{\pi,k,\ell}]_{\ell,k=1}^N$, $p = \text{vec}(P_p)$, $(\bar{A} \otimes \bar{A} + C_A)p = p$, and $p^\top \mathbf{1}_{N^2} = 1$.

^c Mean convergence rates. The matrices $\{\bar{\mathcal{B}}, \bar{B}\}$ are given by (34) from Part II [3] and (46) in this part. Moreover, $\rho_o \triangleq 1 - \lambda_{\min}(H) = 1 - O(\nu)$, where H is given by (84) from Part II [3].

^d Mean-Square convergence rates. The matrices $\{\bar{\mathcal{F}}_{\text{sync}}, \bar{\mathcal{F}}_{\text{async}}, F_{\text{sync}}, F_{\text{async}}\}$ are given by (90), (89), (94), and (51), respectively.

^e Mean-Square-Deviations. The matrices $\{R_{\text{sync}}, R_{\text{async}}\}$ are given by (98) and (63), respectively, and γ_o is given by (70) of Part II [3]. Moreover, $R_{\text{async}} - R_{\text{sync}} = O(\nu^2) > 0$.

For the remainder of this part, we continue to use the same symbols, notation, and assumptions from Part I [2] and Part II [3]. Moreover, we focus on presenting the main results and their interpretation in the body of the paper, while delaying the technical derivations and arguments to the appendices.

II. CENTRALIZED BATCH SOLUTION

We first describe and examine the centralized (batch) solution. In order to allow for a fair comparison among the various implementations, we assume that the centralized solution is also running a stochastic-gradient approximation algorithm albeit one that has access to the *entire* set of data at each iteration. Obviously, centralized solutions can be more powerful and run more complex optimization procedures. Our purpose is to examine the various implementations under similar algorithmic structures and complexity.

A. Centralized Solution in Two Forms

We thus consider a scenario where there is a fusion center that regularly collects the data from across the network and is interested in solving the same minimization problem (1) as in Part I [2], namely,

$$\underset{w}{\text{minimize}} \quad J^{\text{glob}}(w) \triangleq \sum_{k=1}^N J_k(w) \quad (6)$$

where the cost functions $\{J_k(w)\}$ satisfy Assumptions 1 and 2 in Part I [2] and each has a unique minimizer at $w^o \in \mathbb{C}^M$. The fusion center seeks the optimal solution w^o of (6) by running an *asynchronous* stochastic gradient batch algorithm of the following form (later in (66) we consider a synchronous version of this batch solution):

$$\mathbf{w}_{c,i} = \mathbf{w}_{c,i-1} - \sum_{k=1}^N \pi_k(i) \mu_k(i) \widehat{\nabla_{w^*} J_k}(\mathbf{w}_{c,i-1}) \quad (7)$$

where $\mathbf{w}_{c,i}$ denotes the iterate at time i , the $\{\pi_k(i)\}$ are nonnegative convex fusion coefficients such that

$$\sum_{k=1}^N \pi_k(i) = 1, \quad \pi_k(i) \geq 0 \quad (8)$$

for all $i \geq 0$, and the $\{\mu_k(i)\}$ are the random step-sizes.

We will describe later in (66) the centralized implementation for the synchronous batch solution. In that implementation, all agents transmit their data to the fusion center. In contrast, the implementation in (7) allows the transmission of data from agents to occur in an asynchronous manner. Specifically, we use *random* step-sizes $\{\mu_k(i)\}$ in (66) to account for random activity by the agents, which may be caused by random data arrival times or by some power saving strategies that turn agents on and off randomly.

We also use *random* fusion coefficients $\{\pi_k(i)\}$ to model the random status of the communication links connecting the agents to the fusion center. This source of randomness may be caused by random fading effects over the communication channels or by random data feeding/fetching strategies. Therefore, the implementation in (7) is able to accommodate various forms of asynchronous events arising from practical scenarios, and is a useful extension of the classical batch solution in (66).

It is worth noting that the centralized (batch) algorithm (7) admits a decentralized, though not fully-distributed, implementation of the following form:

$$\psi_{k,i} = \mathbf{w}_{c,i-1} - \mu_k(i) \widehat{\nabla_{w^*} J_k}(\mathbf{w}_{c,i-1}) \quad (\text{adaptation}) \quad (9a)$$

$$\mathbf{w}_{c,i} = \sum_{k=1}^N \pi_k(i) \psi_{k,i} \quad (\text{fusion}) \quad (9b)$$

In this description, each agent k uses the local gradient data to calculate the intermediate iterate $\psi_{k,i}$ and feeds its value to a fusion center; the fusion center fuses all intermediate updates $\{\psi_{k,i}\}$ according to (9b) to obtain $\mathbf{w}_{c,i}$ and then forwards the results to all agents. This process repeats itself at every iteration. Implementation (9a)–(9b) is not fully distributed because, for example, all agents require knowledge of the same global iterate $\mathbf{w}_{c,i}$ to perform the adaptation step (9a). Since the one-step centralized implementation (7) and the two-step equivalent (9a)–(9b) represent the same algorithm, we shall use them interchangeably to facilitate the analysis whenever necessary. One advantage of the decentralized representation (9a)–(9b) is that it can be viewed as a distributed solution over *fully*-connected networks [17].

B. Gradient Noise and Asynchronous Models

We assume that the approximate gradient vector $\widehat{\nabla_{w^*} J_k}(\mathbf{w}_{c,i-1})$ in (7) follows the same model described by (18) in Part I [2], namely,

$$\widehat{\nabla_{w^*} J_k}(\mathbf{w}_{c,i-1}) = \nabla_{w^*} J_k(\mathbf{w}_{c,i-1}) + \mathbf{v}_{k,i}(\mathbf{w}_{c,i-1}) \quad (10)$$

where the first term on the RHS is the true gradient and the second term models the uncertainty about the true gradient. We continue to assume that the gradient noise $\mathbf{v}_{k,i}(\mathbf{w}_{c,i-1})$ satisfies Assumption 1 from Part II [3].

From Assumption 1 of Part II [3], the conditional moments of $\mathbf{v}_{k,i}(\mathbf{w}_{c,i-1})$ satisfy

$$\mathbb{E}[\mathbf{v}_{k,i}(\mathbf{w}_{c,i-1}) | \mathbb{F}_{i-1}] = 0 \quad (11)$$

$$\mathbb{E}[\|\mathbf{v}_{k,i}(\mathbf{w}_{c,i-1})\|^4 | \mathbb{F}_{i-1}] \leq \alpha^2 \|\underline{\mathbf{w}}^o - \underline{\mathbf{w}}_{k,i-1}\|^4 + 4\sigma_v^4 \quad (12)$$

where a factor of 4 appeared due to the transform $\mathbb{T}(\cdot)$ from (4) of Part I [2].

To facilitate the comparison in the sequel, we further assume the following asynchronous model for the centralized batch solution (7):

- 1) The random step-sizes $\{\boldsymbol{\mu}_k(i)\}$ satisfy the same properties as the asynchronous model for distributed diffusion networks described in Section III-B of Part I [2]. In particular, the first and second-order moments of $\{\boldsymbol{\mu}_k(i)\}$ are constant and denoted by

$$\bar{\boldsymbol{\mu}}_k \triangleq \mathbb{E}[\boldsymbol{\mu}_k(i)] \quad (13)$$

$$c_{\boldsymbol{\mu},k,\ell} \triangleq \mathbb{E}[(\boldsymbol{\mu}_k(i) - \bar{\boldsymbol{\mu}}_k)(\boldsymbol{\mu}_\ell(i) - \bar{\boldsymbol{\mu}}_\ell)] \quad (14)$$

for all k, ℓ , and $i \geq 0$, where the values of these moments are the same as those in (34) and (37) from Part I [2].

- 2) The random fusion coefficients $\{\boldsymbol{\pi}_k(i)\}$ satisfy condition (8) at every iteration i . Moreover, the first and second-order moments of $\{\boldsymbol{\pi}_k(i)\}$ are denoted by

$$\bar{\boldsymbol{\pi}}_k \triangleq \mathbb{E}[\boldsymbol{\pi}_k(i)] \quad (15)$$

$$c_{\boldsymbol{\pi},k,\ell} \triangleq \mathbb{E}[(\boldsymbol{\pi}_k(i) - \bar{\boldsymbol{\pi}}_k)(\boldsymbol{\pi}_\ell(i) - \bar{\boldsymbol{\pi}}_\ell)] \quad (16)$$

for all k, ℓ , and $i \geq 0$.

- 3) The random parameters $\{\boldsymbol{\mu}_k(i)\}$ and $\{\boldsymbol{\pi}_k(i)\}$ are mutually-independent and independent of any other random variable.

We collect the fusion coefficients into the vector:

$$\boldsymbol{\pi}_i \triangleq \text{col}\{\boldsymbol{\pi}_1(i), \boldsymbol{\pi}_2(i), \dots, \boldsymbol{\pi}_N(i)\} \quad (17)$$

Then, condition (8) implies that $\boldsymbol{\pi}_i^\top \mathbf{1}_N = 1$. By (15) and (16), the mean and covariance matrix of $\boldsymbol{\pi}_i$ are given by

$$\bar{\boldsymbol{\pi}} \triangleq \mathbb{E}(\boldsymbol{\pi}_i) = \text{col}\{\bar{\boldsymbol{\pi}}_1, \bar{\boldsymbol{\pi}}_2, \dots, \bar{\boldsymbol{\pi}}_N\} \quad (18)$$

$$C_{\boldsymbol{\pi}} \triangleq \mathbb{E}[(\boldsymbol{\pi}_i - \bar{\boldsymbol{\pi}})(\boldsymbol{\pi}_i - \bar{\boldsymbol{\pi}})^\top] = \begin{bmatrix} c_{\boldsymbol{\pi},1,1} & \dots & c_{\boldsymbol{\pi},1,N} \\ \vdots & \ddots & \vdots \\ c_{\boldsymbol{\pi},N,1} & \dots & c_{\boldsymbol{\pi},N,N} \end{bmatrix} \quad (19)$$

Lemma 1 (Properties of moments of $\{\boldsymbol{\pi}_k(i)\}$): The first and second-order moments of $\{\boldsymbol{\pi}_k(i)\}$ defined in (15) and (16) satisfy

$$\sum_{k=1}^N \bar{\boldsymbol{\pi}}_k = \mathbf{1}, \quad \bar{\boldsymbol{\pi}}_k \geq 0, \quad \sum_{k=1}^N c_{\boldsymbol{\pi},k,\ell} = 0, \quad \sum_{\ell=1}^N c_{\boldsymbol{\pi},k,\ell} = 0 \quad (20)$$

for any k and ℓ .

Proof: Using (18) and (19) and the fact that C_π is symmetric, conditions (20) require

$$\bar{\pi}^\top \mathbf{1}_N = 1, \quad C_\pi \mathbf{1}_N = 0 \quad (21)$$

The first equation in (21) is straightforward from (8). The second condition in (21) is true because

$$C_\pi \mathbf{1}_N = [\mathbb{E}(\boldsymbol{\pi}_i \boldsymbol{\pi}_i^\top | \mathbf{w}_{c,i-1}) - \bar{\pi} \bar{\pi}^\top] \mathbf{1}_N = 0 \quad (22)$$

where we used the fact that $\boldsymbol{\pi}_i^\top \mathbf{1}_N = 1$ and $\bar{\pi}^\top \mathbf{1}_N = 1$. \blacksquare

We next examine the stability and steady-state performance of the asynchronous batch algorithm (7), and then compare its performance with that of the asynchronous distributed diffusion strategy.

III. PERFORMANCE OF THE CENTRALIZED SOLUTION

Following an argument similar to that given in Section V of Part I [2], we can derive from (9a)–(9b) the following error recursion for the asynchronous centralized implementation:

$$\tilde{\boldsymbol{\psi}}_{k,i} = [I_{2M} - \boldsymbol{\mu}_k(i) \mathbf{H}_{k,i-1}] \tilde{\mathbf{w}}_{c,i-1} + \mathbf{s}_{k,i} \quad (23a)$$

$$\tilde{\mathbf{w}}_{c,i} = \sum_{k=1}^N \boldsymbol{\pi}_k(i) \tilde{\boldsymbol{\psi}}_{k,i} \quad (23b)$$

where

$$\tilde{\mathbf{w}}_{c,i} \triangleq \mathbb{T}(\tilde{\mathbf{w}}_{c,i}) \quad (24)$$

$$\tilde{\boldsymbol{\psi}}_{k,i} \triangleq \mathbb{T}(\tilde{\boldsymbol{\psi}}_{k,i}) \quad (25)$$

$$\mathbf{v}_{k,i}(\mathbf{w}_{c,i-1}) \triangleq \mathbb{T}(\mathbf{v}_{k,i}(\mathbf{w}_{c,i-1})) \quad (26)$$

and the mapping $\mathbb{T}(\cdot)$ is from (4) in Part I [2]. Moreover,

$$\mathbf{H}_{k,i-1} \triangleq \int_0^1 \nabla_{\mathbf{w} \mathbf{w}^*}^2 J_k(\mathbf{w}^o - t \tilde{\mathbf{w}}_{c,i-1}) dt \quad (27)$$

$$\mathbf{s}_{k,i} \triangleq \boldsymbol{\mu}_k(i) \mathbf{v}_{k,i}(\mathbf{w}_{c,i-1}) \quad (28)$$

We can merge (23a) and (23b) to find that the error dynamics of (7) evolves according to the following recursion:

$$\tilde{\mathbf{w}}_{c,i} = \left[I_{2M} - \sum_{k=1}^N \boldsymbol{\pi}_k(i) \boldsymbol{\mu}_k(i) \mathbf{H}_{k,i-1} \right] \tilde{\mathbf{w}}_{c,i-1} + \mathbf{s}_i \quad (29)$$

where

$$\mathbf{s}_i \triangleq \sum_{k=1}^N \boldsymbol{\pi}_k(i) \mathbf{s}_{k,i} = \sum_{k=1}^N \boldsymbol{\pi}_k(i) \boldsymbol{\mu}_k(i) \mathbf{v}_{k,i}(\mathbf{w}_{c,i-1}) \quad (30)$$

A. Mean-Square and Mean-Fourth-Order Stability

To maintain consistency with the notation used in Parts I [2] and II [3], we shall employ the same auxiliary quantities in these parts for the centralized batch solution (7) with minor adjustments whenever necessary. For example, the error quantity $\tilde{\mathbf{w}}_{k,i}$ used before in Parts I [2] and II [3] for the error vector at agent k at time i in the distributed implementation is now replaced by $\tilde{\mathbf{w}}_{c,i}$, with a subscript c , for the error vector of the centralized solution at time i . Thus, we let

$$\epsilon^2(i) \triangleq \mathbb{E} \|\tilde{\mathbf{w}}_{c,i}\|^2 = \frac{1}{2} \mathbb{E} \|\tilde{\mathbf{w}}_{c,i}\|^2 \quad (31)$$

denote the MSD for the centralized solution $\tilde{\mathbf{w}}_{c,i}$.

Theorem 1 (Mean-square stability): The mean-square stability of the asynchronous centralized implementation (7) reduces to studying the convergence of the recursive inequality:

$$\epsilon^2(i) \leq \beta \cdot \epsilon^2(i-1) + \theta \sigma_v^2 \quad (32)$$

where the parameters $\{\beta, \theta, \sigma_v^2\}$ are from (90), (91), and (25) in Part I [2], respectively. The model (32) is stable if condition

$$\boxed{\frac{\bar{\mu}_k^{(2)}}{\bar{\mu}_k^{(1)}} < \frac{\lambda_{k,\min}}{\lambda_{k,\max}^2 + \alpha}} \quad (33)$$

holds for all k , where the parameters $\{\lambda_{k,\min}, \lambda_{k,\max}, \alpha\}$ are from Assumptions 2 and 3 of Part I [2], respectively. When condition (33) holds, an upper bound on the steady-state MSD is given by

$$\boxed{\limsup_{i \rightarrow \infty} \mathbb{E} \|\tilde{\mathbf{w}}_{c,i}\|^2 \leq b \cdot \nu} \quad (34)$$

where ν is given by (3) and b is a constant defined by (95) from Part I [2].

Proof: Since the centralized solution (7), or, equivalently, (9a)–(9b), can be viewed as a distributed solution over *fully*-connected networks [17], Theorem 1 from Part I [2] can be applied directly. The result then follows from the fact that $\nu_o \leq \nu$ by (107) in Part I [2]. ■

Comparing the above result to Theorem 1 in Part I [2], we observe that the mean-square stability of the centralized solution (7) and the distributed asynchronous solution (39a)–(39b) from Part I [2] is governed by the same model (32). Therefore, the same condition (33) guarantees the stability for both strategies and leads to the same MSD bound (34).

Theorem 2 (Stability of fourth-order error moment): If

$$\boxed{\frac{\sqrt{\bar{\mu}_k^{(4)}}}{\bar{\mu}_k^{(1)}} < \frac{\lambda_{k,\min}}{3\lambda_{k,\max}^2 + 4\alpha}} \quad (35)$$

holds for all k , then the fourth-order moment of the error $\tilde{\mathbf{w}}_{c,i}$ is asymptotically bounded by

$$\boxed{\limsup_{i \rightarrow \infty} \mathbb{E} \|\tilde{\mathbf{w}}_{c,i}\|^4 \leq b_4^2 \cdot \nu^2} \quad (36)$$

where the parameter ν is given by (3), and b_4 is a constant defined by (105) of Part I [2].

Proof: This result follows from Theorem 2 of Part I [2] because the centralized solution (7), or, equivalently, (9a)–(9b), can be viewed as a distributed solution over *fully*-connected networks [17]. ■

An alternative method to investigate the stability conditions for the centralized solution (7) is to view it as a stochastic gradient descent iteration for a *standalone* agent (i.e., a singleton network with $N = 1$) [14]–[16].

B. Long Term Error Dynamics

Using an argument similar to the one in Section II-A from Part II [3], the original The original error recursion (29) can be rewritten as

$$\tilde{\mathbf{w}}_{c,i} = \left[I_{2M} - \sum_{k=1}^N \pi_k(i) \boldsymbol{\mu}_k(i) H_k \right] \tilde{\mathbf{w}}_{c,i-1} + \underline{\mathbf{s}}_i + \mathbf{d}_i \quad (37)$$

where

$$\mathbf{d}_i \triangleq \sum_{k=1}^N \pi_k(i) \boldsymbol{\mu}_k(i) (H_k - \mathbf{H}_{k,i-1}) \tilde{\mathbf{w}}_{c,i-1} \quad (38)$$

Then, under condition (35),

$$\limsup_{i \rightarrow \infty} \mathbb{E} \|\mathbf{d}_i\|^2 \leq O(\nu^4) \quad (39)$$

where ν is given by (3).

Assumption 1 (Small step-sizes): The parameter ν from (3) is sufficiently small such that

$$\nu < \min_k \frac{\lambda_{k,\min}}{3\lambda_{k,\max}^2 + 4\alpha} < 1 \quad (40)$$

■

Under Assumption 1, condition (35) holds. Let

$$\mathbf{B}_i \triangleq \sum_{k=1}^N \pi_k(i) \mathbf{D}_{k,i} \quad (41)$$

$$\mathbf{D}_{k,i} \triangleq I_{2M} - \boldsymbol{\mu}_k(i) H_k \quad (42)$$

where \mathbf{B}_i is Hermitian positive semi-definite. Since we are interested in examining the asymptotic performance of the asynchronous batch solution, we can again call upon the same argument from Section

II-A of Part II [3] and use result (39) to conclude that we can assess the performance of (37) by working with the following *long-term* model, which holds for large enough i :

$$\tilde{\mathbf{w}}'_{c,i} = \mathbf{B}_i \cdot \tilde{\mathbf{w}}'_{c,i-1} + \underline{\mathbf{s}}_i, \quad i \gg 1 \quad (43)$$

In model (43), we ignored the $O(\nu^2)$ term \mathbf{d}_i according to (39), and we are using $\mathbf{w}'_{c,i}$ to denote the estimate obtained from this long-term model. Note that the driving noise term $\underline{\mathbf{s}}_i$ in (43) is extraneous and imported from the original error recursion (29).

Theorem 3 (Bounded mean-square gap): Under Assumption 1, the mean-square gap from the original error recursion (29) to the long-term model (43) is asymptotically bounded by

$$\limsup_{i \rightarrow \infty} \mathbb{E} \|\tilde{\mathbf{w}}_{c,i} - \tilde{\mathbf{w}}'_{c,i}\|^2 \leq O(\nu^2) \quad (44)$$

where ν is given by (3).

Proof: This result follows from Theorem 1 of Part II [3] since the centralized solution (7) can be viewed as a distributed solution over *fully-connected* networks [17]. ■

C. Mean Error Recursion

By taking the expectation of both sides of (43), and using the fact that $\mathbb{E}(\underline{\mathbf{s}}_i) = 0$, we conclude that the mean error satisfies the recursion:

$$\mathbb{E} \tilde{\mathbf{w}}'_{c,i} = \bar{\mathbf{B}} \cdot \mathbb{E} \tilde{\mathbf{w}}'_{c,i-1} \quad (45)$$

for large enough i , where

$$\bar{\mathbf{B}} \triangleq \mathbb{E}(\mathbf{B}_i) = \sum_{k=1}^N \bar{\pi}_k \bar{\mathbf{D}}_k \quad (46)$$

$$\bar{\mathbf{D}}_k \triangleq \mathbb{E}(\mathbf{D}_{k,i}) = \mathbf{I}_{2M} - \bar{\mu}_k \mathbf{H}_k \quad (47)$$

The convergence of recursion (45) requires the stability of $\bar{\mathbf{B}}$. It is easy to verify that $\{\bar{\mathbf{B}}, \bar{\mathbf{D}}_k\}$ are Hermitian. Using (20) and Jensen's inequality, we get from (46) that $\rho(\bar{\mathbf{B}}) \leq \max_k \rho(\bar{\mathbf{D}}_k)$. As we showed in (45) from Part II [2], if condition (33) holds, then $\rho(\bar{\mathbf{D}}_k) < 1$ for all k . Therefore, it follows from Assumption 1 that

$$\lim_{i \rightarrow \infty} \mathbb{E} \tilde{\mathbf{w}}'_{c,i} = 0 \quad (48)$$

which implies that the long-term model (43) is the asymptotically centered version of the original error recursion (29).

D. Error Covariance Recursion

Let \mathbb{F}_{i-1} denote the filtration that represents all information available up to iteration $i - 1$. Then we deduce from (43) that for large enough i :

$$\mathbb{E}(\tilde{\mathbf{w}}'_{c,i} \tilde{\mathbf{w}}'^{*}_{c,i} | \mathbb{F}_{i-1}) = \mathbb{E}(\mathbf{B}_i \tilde{\mathbf{w}}'_{c,i-1} \tilde{\mathbf{w}}'^{*}_{c,i-1} \mathbf{B}_i | \mathbb{F}_{i-1}) + \mathbb{E}(\mathbf{s}_i \mathbf{s}_i^* | \mathbb{F}_{i-1}) \quad (49)$$

where the cross terms that involve \mathbf{s}_i disappear because $\mathbb{E}(\mathbf{s}_i^* \mathbf{B}_i \tilde{\mathbf{w}}'_{c,i-1} | \mathbb{F}_{i-1}) = 0$ by the gradient noise model from Assumption 1 of Part II [3]. Vectorizing both sides of (49) and taking expectation, we obtain

$$\mathbb{E}[(\tilde{\mathbf{w}}'^{*}_{c,i})^\top \otimes \tilde{\mathbf{w}}'_{c,i}] = F_c \cdot \mathbb{E}[(\tilde{\mathbf{w}}'^{*}_{c,i-1})^\top \otimes \tilde{\mathbf{w}}'_{c,i-1}] + y_{c,i} \quad (50)$$

where

$$F_c \triangleq \mathbb{E}[\mathbf{B}_i^\top \otimes \mathbf{B}_i] \quad (51)$$

$$y_{c,i} \triangleq \mathbb{E}[(\mathbf{s}_i^*)^\top \otimes \mathbf{s}_i] \quad (52)$$

Let further

$$H_c \triangleq \sum_{k=1}^N \bar{\pi}_k \bar{\mu}_k H_k = O(\nu) \quad (53)$$

where $\{H_k\}$ are from (14) of Part II [3].

Lemma 2 (Properties of F_c): The matrix F_c defined by (51) is Hermitian and can be expressed as

$$F_c = \sum_{\ell=1}^N \sum_{k=1}^N (\bar{\pi}_\ell \bar{\pi}_k + c_{\pi,\ell,k}) (\bar{D}_\ell^\top \otimes \bar{D}_k + c_{\mu,\ell,k} H_\ell^\top \otimes H_k) \quad (54)$$

If condition (33) holds, then F_c is stable and

$$\rho(F_c) = [1 - \lambda_{\min}(H_c)]^2 + O(\nu^2) \quad (55)$$

where H_c is given by (53), and $[1 - \lambda_{\min}(H_c)]^2 = 1 - O(\nu)$ under Assumption 1. Moreover,

$$\|(I_{4M^2} - F_c)^{-1}\| = O(\nu^{-1}) \quad (56)$$

Proof: See Appendix A. ■

Theorem 4 (Error covariance recursion): For sufficiently large i , the vectorized error covariance for the long-term model (43) satisfies the following relation:

$$z_{c,i} = F_c \cdot z_{c,i-1} + y_{c,i}, \quad i \gg 1 \quad (57)$$

where F_c and $y_{c,i}$ are from (51) and (52), respectively, and

$$z_{c,i} \triangleq \mathbb{E}[(\tilde{\mathbf{w}}'^{*}_{c,i})^\top \otimes \tilde{\mathbf{w}}'_{c,i}] \quad (58)$$

Recursion (57) is convergent if condition (33) holds, and its convergence rate is dominated by $[1 - \lambda_{\min}(H_c)]^2 = 1 - O(\nu)$ under Assumption 1.

Proof: Equation (57) follows from (50). Recursion (57) converges if, and only if, the matrix F_c is stable. By Lemma 2, we know that $\rho(F_c) < 1$ if condition (33) holds and, moreover, the convergence rate of recursion (57) is determined by $\rho(F_c) = [1 - \lambda_{\min}(H_c)]^2 + O(\nu^2)$. ■

E. Steady-State MSD

At steady-state as $i \rightarrow \infty$, we get from (56) and (57) that

$$\begin{aligned} z_{c,\infty} &\triangleq \text{vec} \left(\lim_{i \rightarrow \infty} \mathbb{E} \tilde{\mathbf{w}}'_{c,i} \tilde{\mathbf{w}}'^*_{c,i} \right) \\ &= (I_{4M^2} - F_c)^{-1} \cdot \lim_{i \rightarrow \infty} y_{c,i} \end{aligned} \quad (59)$$

Using $z_{c,\infty}$, we can determine the value of any steady-state weighted mean-square-error metric for the long-term model (43) as follows:

$$\begin{aligned} \lim_{i \rightarrow \infty} \mathbb{E} \|\tilde{\mathbf{w}}'_{c,i}\|_{\Sigma}^2 &= \frac{1}{2} \lim_{i \rightarrow \infty} \text{Tr}[\mathbb{E}(\tilde{\mathbf{w}}'_{c,i} \tilde{\mathbf{w}}'^*_{c,i}) \Sigma] \\ &= \frac{1}{2} z_{c,\infty}^* \text{vec}(\Sigma) \end{aligned} \quad (60)$$

where we used the fact that $\text{Tr}(AB) = [\text{vec}(A^*)]^* \text{vec}(B)$, and Σ is an arbitrary Hermitian positive semi-definite weighting matrix. The steady-state MSD for the original error recursion (37) is defined by

$$\text{MSD}^{\text{cent}} \triangleq \lim_{i \rightarrow \infty} \mathbb{E} \|\tilde{\mathbf{w}}_{c,i}\|^2 = \lim_{i \rightarrow \infty} \frac{1}{2} \mathbb{E} \|\tilde{\mathbf{w}}_{c,i}\|^2 \quad (61)$$

Therefore, by setting $\Sigma = I_{2M}$ in (60) and using Theorem 3, it is easy to verify by following an argument similar to the proof of Theorem 3 from Part II [3] that

$$\text{MSD}^{\text{cent}} = \lim_{i \rightarrow \infty} \mathbb{E} \|\tilde{\mathbf{w}}'_{c,i}\|^2 + O(\nu^{3/2}) \quad (62)$$

Introduce

$$R_c \triangleq \sum_{k=1}^N (\bar{\pi}_k^2 + c_{\pi,k,k}) (\bar{\mu}_k^2 + c_{\mu,k,k}) R_k = O(\nu^2) \quad (63)$$

where $\{R_k\}$ are from (10) of Part II [3]. Then, using (60) and (62), we arrive the following result.

Theorem 5 (Steady-state MSD): The steady-state MSD for the asynchronous centralized (batch) solution (7) is given by

$$\text{MSD}^{\text{cent}} = \frac{1}{2} [\text{vec}(R_c)]^* (I_{4M^2} - F_c)^{-1} \text{vec}(I_{2M}) + O(\nu^{1+\gamma_o}) \quad (64)$$

where $0 < \gamma_o \leq 1/2$ is from (70) of Part II [3]. Expression (64) can be further reworked to yield

$$\text{MSD}^{\text{cent}} = \frac{1}{4} \text{Tr}(H_c^{-1} R_c) + O(\nu^{1+\gamma_o}) \quad (65)$$

where the first term on the RHS is in the order of ν and therefore dominates the $O(\nu^{1+\gamma_o})$ term under Assumption 1.

Proof: See Appendix B. ■

F. Results for the Synchronous Centralized Solution

We may also consider a synchronous centralized (batch) implementation for solving the same problem (6). It would take the following form:

$$\mathbf{w}_{c,i} = \mathbf{w}_{c,i-1} - \sum_{k=1}^N \pi_k \mu_k \widehat{\nabla_{w^*} J_k}(\mathbf{w}_{c,i-1}) \quad (66)$$

where the $\{\mu_k\}$ are now deterministic nonnegative step-sizes and the $\{\pi_k\}$ are nonnegative fusion coefficients that satisfy $\sum_{k=1}^N \pi_k = 1$. The synchronous batch solution can be viewed as a special case of the asynchronous batch solution (7) when the random step-sizes and fusion coefficients assume constant values. If the covariances $\{c_{\mu,k,k}\}$ and $\{c_{\pi,k,k}\}$ are set to zero, then the asynchronous solution (7) will reduce into a synchronous solution that employs the constant parameters $\{\bar{\mu}_k\}$ and $\{\bar{\pi}_k\}$. The previous stability and performance results can be specialized to the synchronous batch implementation under these conditions.

It is easy to verify that the mean error recursion for the synchronous solution with parameters $\{\bar{\mu}_k\}$ and $\{\bar{\pi}_k\}$ is identical to (45). The mean convergence rate for the long-term model is still determined by $\rho(\bar{B})$, where \bar{B} is given by (46). The mean square convergence rate for the long-term model is determined by $\rho(F'_c)$ where

$$F'_c \triangleq \sum_{k=1}^N \sum_{\ell=1}^N \bar{\pi}_\ell \bar{\pi}_k (\bar{D}_\ell^T \otimes \bar{D}_k) \quad (67)$$

It follows that

$$\rho(F'_c) = [1 - \lambda_{\min}(H_c)]^2 = 1 - O(\nu) \quad (68)$$

The steady-state MSD is given by

$$\text{MSD}_{\text{sync}}^{\text{cent}} = \frac{1}{4} \text{Tr}(H_c^{-1} R'_c) + O(\nu^{1+\gamma_o}) \quad (69)$$

where

$$R'_c \triangleq \sum_{k=1}^N \bar{\pi}_k^2 \bar{\mu}_k^2 R_k, \quad \|R'_c\| = O(\nu^2) \quad (70)$$

and $\text{Tr}(H_c^{-1} R'_c) = O(\nu)$.

IV. COMPARISON I: DISTRIBUTED VS. CENTRALIZED STRATEGIES

In this section, we compare the mean-square performance of the distributed diffusion strategy (28a)–(28b) from Part I [2], namely,

$$\psi_{k,i} = \mathbf{w}_{k,i-1} - \mu_k(i) \widehat{\nabla_{w^*} J_k}(\mathbf{w}_{k,i-1}) \quad (71a)$$

$$\mathbf{w}_{k,i} = \sum_{\ell \in \mathcal{N}_{k,i}} \mathbf{a}_{\ell k}(i) \psi_{\ell,i} \quad (71b)$$

with the centralized (batch) solution described by (7). We establish the important conclusion that if the combination matrix is primitive (Assumption 3 in Part II [3]), then the asynchronous network is able to achieve almost the same mean-square performance as the centralized (batch) solution for sufficiently small step-sizes. In other words, diffusion strategies are *efficient* mechanisms to perform continuous adaptation and learning tasks over networks even in the presence of various sources of random failures.

A. Adjusting Relevant Parameters

First, however, we need to describe the conditions that are necessary for a *fair* and meaningful comparison between the distributed and centralized implementations. This is because the two implementations use different parameters. Recall that the agents in the distributed network (71a)–(71b) employ random combination coefficients $\{\mathbf{a}_{\ell k}(i)\}$ to aggregate information from neighborhoods using random step-sizes $\{\mu_k(i)\}$. The random parameters $\{\mathbf{a}_{\ell k}(i), \mu_k(i)\}$ are assumed to satisfy the model described in Section III-B from Part I [2]. On the other hand, the centralized batch solution (7) uses random combination coefficients $\{\pi_k(i)\}$ to fuse the information from all agents in the network, and then performs updates using random step-sizes $\{\mu_k(i)\}$. The random parameters $\{\pi_k(i), \mu_k(i)\}$ are assumed to satisfy the conditions specified in Section II-B of this part. In general, the two sets of random parameters, i.e., $\{\mathbf{a}_{\ell k}(i), \mu_k(i)\}$ for distributed strategies and $\{\pi_k(i), \mu_k(i)\}$ for centralized strategies, are not necessarily related. Therefore, in order to make a meaningful comparison between the distributed and centralized strategies, we need to introduce connections between these two sets of parameters. This is possible because the parameters play similar roles.

From the previous analysis in Section IV of Part II [3], we know that the first and second-order moments of $\{\mathbf{a}_{\ell k}(i), \mu_k(i)\}$ determine the mean-square performance of diffusion networks. Likewise, from the analysis in Section III of this part, we know that the first and second-order moments of $\{\pi_k(i), \mu_k(i)\}$ determine the mean-square performance of centralized solutions. Therefore, it is sufficient to introduce connections between the first and second-order moments of these random parameters. For the random

step-size parameters, we assumed in (34) and (37) from Part I [2] and in (13) and (14) from this part that their first and second-order moments are *constant* and that their values coincide with each other, i.e., $\bar{\mu}_k$ from (34) in Part I [2] coincides with $\bar{\mu}_k$ from (13) in this part, and similarly for $c_{\mu,k,\ell}$. This requirement is obviously reasonable.

The connection that we need to enforce between the moments of the combination coefficients $\{\mathbf{a}_{\ell k}(i)\}$ and $\{\boldsymbol{\pi}_k(i)\}$, while reasonable again, is less straightforward to explain. This is because the $\{\mathbf{a}_{\ell k}(i)\}$ form a random matrix $\mathbf{A}_i = [\mathbf{a}_{\ell k}(i)]_{k,\ell=1}^N$ of size $N \times N$, while the $\{\boldsymbol{\pi}_k(i)\}$ only form a random vector $\boldsymbol{\pi}_i = [\boldsymbol{\pi}_k(i)]_{k=1}^N$ of size $N \times 1$. From the result of Corollary 2 in Part II [3] though, we know that the mean-square performance of the *primitive* diffusion network does not *directly* depend on the moments of \mathbf{A}_i , namely, its mean $\bar{\mathbf{A}}$ and its Kronecker covariance C_A ; instead, the performance depends on the Perron eigenvector (the unique right eigenvector corresponding to the eigenvalue at one for primitive left-stochastic matrices [18], [19]). If, for example, we compare expression (96) from Part II [3] for asynchronous networks with expression (54) from this part, we conclude that it is sufficient to relate the vectors $\{\bar{p}, p\}$ defined in (77) and (78) from Part II [3] to the moments $\{\bar{\pi}_k, c_{\pi,k,\ell}\}$. Since \bar{p} is the Perron eigenvector of the mean matrix $\bar{\mathbf{A}}$, and the $\{\bar{\pi}_k\}$ are the means of $\{\boldsymbol{\pi}_k(i)\}$, we connect them by requiring

$$\bar{\pi}_k \equiv \bar{p}_k \quad (72)$$

for all k , where the $\{\bar{p}_k\}$ are the elements of \bar{p} . Likewise, since p is the Perron eigenvector of the matrix $\bar{\mathbf{A}} \otimes \bar{\mathbf{A}} + C_A = \mathbb{E}(\mathbf{A}_i \otimes \mathbf{A}_i)$, which consists of the second-order moments, and $\{\bar{\pi}_k \bar{\pi}_\ell + c_{\pi,k,\ell} = \mathbb{E}[\boldsymbol{\pi}_k(i) \boldsymbol{\pi}_\ell(i)]\}$ are also the second-order moments, we connect them by requiring

$$\bar{\pi}_k \bar{\pi}_\ell + c_{\pi,k,\ell} \equiv p_{k,\ell} \quad (73)$$

for all k and ℓ , where the $\{p_{k,\ell}\}$ are the elements of p defined after (80) in Part II [3]. When conditions (72) and (73) are satisfied, then the mean-square convergence rates and steady-state MSD for the distributed and centralized solutions become identical. We establish this result in the sequel. Using (18) and (19), conditions (72) and (73) can be rewritten as

$$\bar{\pi} \equiv \bar{p}, \quad C_\pi + \bar{\pi} \bar{\pi}^\top \equiv P_p \quad (74)$$

where

$$P_p = \begin{bmatrix} p_{1,1} & \cdots & p_{1,N} \\ \vdots & \ddots & \vdots \\ p_{N,1} & \cdots & p_{N,N} \end{bmatrix} \quad (75)$$

is the symmetric matrix defined by (80) of Part II [3]. It is worth noting that, since the Perron eigenvectors $p = \text{vec}(P_p)$ and \bar{p} consist of positive entries, the corresponding quantities $\bar{\pi}$ and $C_\pi + \bar{\pi}\bar{\pi}^\top$ must also consist of positive entries — we shall refer to the centralized solutions that satisfy this condition as *primitive* centralized solutions. Clearly, the second requirement in (74) is meaningful only if the difference $P_p - \bar{p}\bar{p}^\top$ results in a symmetric positive semi-definite matrix (and, hence, a covariance matrix) that also satisfies $C_\pi \mathbb{1}_N = 0$.

B. Constructing Primitive Batch Solutions

Before comparing the performance of the centralized and distributed solutions under (74), we first answer the following important inquiry. Given a distributed primitive network with parameters $\{\bar{p}, P_p\}$, is it possible to determine a batch solution with parameters $\{\bar{\pi}, C_\pi\}$ satisfying (74) such that the resulting C_π is a symmetric and positive semi-definite matrix (and, therefore, has the interpretation of a valid covariance matrix)? The answer is in the affirmative as we proceed to explain. The following are auxiliary results in this direction.

Lemma 3 (Positive semi-definite property): The matrix difference $P_p - \bar{p}\bar{p}^\top$ is symmetric positive semi-definite and satisfies $(P_p - \bar{p}\bar{p}^\top) \mathbb{1}_N = 0$ for any \bar{p} and P_p defined by (78) and (80) from Part II [3].

Proof: See Appendix C. ■

Therefore, starting from an asynchronous diffusion network with parameters $\{\bar{p}, P_p\}$, there exists an asynchronous batch solution with valid parameters $\{\bar{\pi}, C_\pi\}$ that satisfy (74). We now explain one way by which a random variable π_i can be constructed with the pre-specified moments $\{\bar{\pi}, C_\pi\}$. We first observe that in view of condition (17), the random variable π_i is actually defined on the probability simplex in $\mathbb{R}^{N \times 1}$ [20, p. 33]:

$$\Delta_N \triangleq \{x \in \mathbb{R}^{N \times 1}; x^\top \mathbb{1}_N = 1, x_k \geq 0, k = 1, \dots, N\} \quad (76)$$

If the moments $\{\bar{\pi}, C_\pi\}$ obtained from (74) satisfy certain conditions, then there are several models in the literature that can be used to generate random vectors $\{\pi_i\}$ according to these moments such as using the Dirichlet distribution [21], the Generalized Dirichlet distribution [22]–[29], the Logistic-Normal distribution [23], [30], [31], or the Generalized inverse Gaussian distribution [24], [32]. Unfortunately, if the conditions for these models are not satisfied, no *closed-form* probabilistic model is available for us to generate random variables on the probability simplex with pre-specified means and covariance matrices.

Nevertheless, inspired by the Markov Chain Monte Carlo (MCMC) method [33], we describe one procedure to construct random variables *indirectly* so that they are able to meet the desired moment

requirements. In a manner similar to the argument used in Appendix C, we introduce a series of fictitious random combination matrices $\{\mathbf{A}'_j; j \geq 1\}$ that satisfy the asynchronous model introduced in Part I [2]. We assume that the $\{\mathbf{A}'_j; j \geq 1\}$ are independently, identically distributed (i.i.d.) random matrices, and they are independent of any other random variable. Then, the mean and Kronecker-covariance matrices of \mathbf{A}'_j for any j are given by \bar{A} and C_A , respectively. We further introduce the random matrix

$$\Phi_{i,t} \triangleq \prod_{j=1}^t \mathbf{A}'_j \quad (77)$$

Similar to (161) and (163), we can verify that

$$\lim_{t \rightarrow \infty} \mathbb{E}(\Phi_{i,t}) = \bar{p} \mathbf{1}_N^\top, \quad \lim_{t \rightarrow \infty} \mathbb{E}(\Phi_{i,t} \otimes \Phi_{i,t}) = p \mathbf{1}_{N^2}^\top \quad (78)$$

Let

$$\phi_i \triangleq \frac{1}{N} \left(\lim_{t \rightarrow \infty} \Phi_{i,t} \right) \mathbf{1}_N \quad (79)$$

Then, the entries of ϕ_i are nonnegative since the entries of $\Phi_{i,t}$ are nonnegative. Using (77) and (79), we have

$$\mathbf{1}_N^\top \phi_i = \frac{1}{N} \lim_{t \rightarrow \infty} \mathbf{1}_N^\top \left(\prod_{j=1}^t \mathbf{A}'_j \right) \mathbf{1}_N = 1 \quad (80)$$

since each \mathbf{A}'_j is left-stochastic. Therefore, ϕ_i is a random variable defined on the probability simplex Δ_N . By using (78) and the fact that $\mathbf{1}_N \otimes \mathbf{1}_N = \mathbf{1}_{N^2}$, we have

$$\mathbb{E}(\phi_i) = \frac{1}{N} (\bar{p} \cdot \mathbf{1}_N^\top) \mathbf{1}_N = \bar{p} \quad (81)$$

$$\mathbb{E}(\phi_i \otimes \phi_i) = \frac{1}{N^2} (p \cdot \mathbf{1}_{N^2}^\top) \mathbf{1}_{N^2} = p \quad (82)$$

Therefore,

$$\mathbb{E}(\phi_i) = \bar{p}, \quad \text{Cov}(\phi_i) = P_p - \bar{p} \bar{p}^\top \quad (83)$$

where $P_p = \text{unvec}(p)$. In this way, we have been able to construct a random variable ϕ_i whose support is the probability simplex Δ_N and whose mean vector and covariance matrix match the specification. The random variable ϕ_i can then be used by the asynchronous centralized solution at time i , which would then enable a meaningful comparison with the asynchronous distributed solution.

Although unnecessary for our development, it is instructive to pose the converse question: Given a *primitive* batch solution with parameters $\{\bar{\pi}, C_\pi\}$, is it always possible to determine a distributed solution with parameters $\{\bar{p}, P_p\}$ satisfying (74) such that these parameters have the properties of Perron eigenvectors? In other words, given a primitive centralized solution, is it possible to determine a distributed

solution on a *partially*-connected network (otherwise the problem is trivial since fully-connected networks are equivalent to centralized solutions [17]) with equivalent performance levels? The answer to this question remains open. The challenge stems from the fact mentioned earlier that, in general, there is no systematic solution to generate distributions on the probability simplex with pre-specified first and second-order moments. The method of moments [34], which is an iterative solution, does not generally guarantee convergence and therefore, cannot ensure that a satisfactory distribution can be generated eventually.

C. Comparing Performance

From the mean error recursion in (42) of Part II [3], the mean convergence rate for the long-term model of the distributed diffusion strategy is determined by $\rho(\bar{\mathcal{B}})$, where $\bar{\mathcal{B}}$ is defined by (34) of Part II [3]. From the mean error recursion (45) in this part, the mean convergence rate for the long-term model of the centralized batch solution is determined by $\rho(\bar{B})$, where \bar{B} is given by (46).

Lemma 4 (Matching mean convergence rates): The mean convergence rates for the asynchronous distributed strategy and the centralized batch solution are almost the same. Specifically, it holds that

$$|\rho(\bar{\mathcal{B}}) - \rho(\bar{B})| \leq O(\nu^{1+1/N}) \quad (84)$$

where $\rho(\bar{\mathcal{B}})$ and $\rho(\bar{B})$ are of the order of $1 - O(\nu)$.

Proof: See Appendix D. ■

Likewise, from Theorem 2 of Part II [3], the mean-square convergence rate of the distributed diffusion strategy for large enough i is determined by $\rho(\mathcal{F})$, where \mathcal{F} is from (51) of Part II [3]. From Theorem 4 of this part, the mean-square convergence rate of the centralized (batch) solution is determined by $\rho(F_c)$, where F_c is from (51).

Lemma 5 (Matching mean-square convergence rates): The mean-square convergence rates for the asynchronous distributed strategy and the centralized batch solution are almost the same. Specifically, it holds that

$$|\rho(\mathcal{F}) - \rho(F_c)| \leq O(\nu^{1+1/N^2}) \quad (85)$$

where $\rho(\mathcal{F})$ and $\rho(F_c)$ are of the order of $1 - O(\nu)$.

Proof: From (51) and (73), it is easy to verify that $F_c = F$, where F is from (82) of Part II [3]. Using Lemmas 4 and 5 from Part II [3] then completes the proof. ■

The steady-state network MSD for the distributed diffusion strategy is given by (96) of Part II [3]:

$$\text{MSD}^{\text{dist}} = \frac{1}{4} \text{Tr}(H^{-1}R) + O(\nu^{1+\gamma_o}) \quad (86)$$

for some $0 < \gamma_o \leq 1/2$ given by (70) of Part II [3]. The steady-state MSD for the centralized batch solution is given by (65).

Lemma 6 (Matching MSD performance): At steady-state, the network MSD for the asynchronous distributed strategy and the MSD for the centralized batch solution are close to each other. Specifically, we have

$$|\text{MSD}^{\text{dist}} - \text{MSD}^{\text{cent}}| \leq O(\nu^{1+\gamma_o}) \quad (87)$$

where both MSD^{dist} and MSD^{cent} are in the order of ν .

Proof: From (53), (63), and (73), it is easy to verify that $H_c = H$ and $R_c = R$, where $\{H, R\}$ are given by (84) and (88) of Part II [3]. Using (65) and (86) then completes the proof. ■

V. COMPARISON II: ASYNCHRONOUS VS. SYNCHRONOUS NETWORKS

Synchronous diffusion networks run (16a)–(16b) from Part I [2], namely,

$$\psi_{k,i} = \mathbf{w}_{k,i-1} - \mu_k \widehat{\nabla_{w^*} J_k}(\mathbf{w}_{k,i-1}) \quad (\text{adaptation}) \quad (88a)$$

$$\mathbf{w}_{k,i} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \psi_{\ell,i} \quad (\text{combination}) \quad (88b)$$

These networks can be viewed as a special case of asynchronous networks running (71a)–(71b) when the random step-sizes and combination coefficients assume constant values. If we set the covariances $\{c_{\mu,k,k}\}$ and $\{c_{a,\ell k,\ell k}\}$ to zero, then the asynchronous network (71a)–(71b) will reduce to the synchronous network (88a)–(88b) with the parameters $\{\mu_k, a_{\ell k}\}$ replaced by $\{\bar{\mu}_k, \bar{a}_{\ell k}\}$. We can therefore specialize the results obtained for asynchronous networks to the synchronous case by using $\{\bar{\mu}_k\}$ and $\{\bar{a}_{\ell k}\}$ and assuming $c_{\mu,k,k} = 0$ and $c_{a,\ell k,\ell k} = 0$ for all k and ℓ . For example, it is easy to verify that the mean error recursion of the long term model for the synchronous solution with $\{\bar{\mu}_k\}$ and $\{\bar{a}_{\ell k}\}$ is identical to (42) from Part II [3] for the asynchronous solution.

Under Assumption 1, the asynchronous network with the random parameters $\{\boldsymbol{\mu}_k(i)\}$ and $\{\mathbf{a}_{\ell k}(i)\}$ and the synchronous network with the constant parameters $\{\bar{\mu}_k\}$ and $\{\bar{a}_{\ell k}\}$ have similar mean-square convergence rates for large i , but the steady-state MSD performance of the former is larger than that of the latter by a small amount. This result is established as follows. From Theorem 2 in Part II [3], the mean-square convergence rate for the asynchronous network with large i is determined by $\rho(\mathcal{F}_{\text{async}})$ where

$$\mathcal{F}_{\text{async}} = \mathbb{E}(\mathbf{B}_i^{\text{T}} \otimes_b \mathbf{B}_i^*) \quad (89)$$

and \mathcal{B}_i is given by (28) of Part II [3]. We are adding the subscript “async” to quantities that are related to asynchronous networks. Correspondingly, the mean-square convergence rate for the synchronous network with the constant parameters $\{\bar{\mu}_k\}$ and $\{\bar{a}_{\ell k}\}$ will be determined by $\rho(\mathcal{F}_{\text{sync}})$ where

$$\mathcal{F}_{\text{sync}} \triangleq \bar{\mathcal{B}}^\top \otimes_b \bar{\mathcal{B}}^* \quad (90)$$

and $\bar{\mathcal{B}}$ is given by (34) of Part II [3].

Lemma 7 (Matching mean-square convergence rates): For large i , the mean-square convergence rate of the asynchronous diffusion strategy is close to that of the synchronous diffusion strategy:

$$|\rho(\mathcal{F}_{\text{async}}) - \rho(\mathcal{F}_{\text{sync}})| = O(\nu^{1+1/N^2}) \quad (91)$$

where $\rho(\mathcal{F}_{\text{async}})$ and $\rho(\mathcal{F}_{\text{sync}})$ are both dominated by $[1 - \lambda_{\min}(H)]^2 = 1 - O(\nu)$ for small ν by Assumption 1.

Proof: By Lemma 5 of Part II [3], we have

$$\rho(\mathcal{F}_{\text{async}}) = \rho(F_{\text{async}}) + O(\nu^{1+1/N^2}) \quad (92)$$

where F_{async} is given by (82) of Part II [3]. Correspondingly, we will also have

$$\rho(\mathcal{F}_{\text{sync}}) = \rho(F_{\text{sync}}) + O(\nu^{1+1/N^2}) \quad (93)$$

where F_{sync} is given by

$$F_{\text{sync}} = \sum_{k=1}^N \sum_{\ell=1}^N \bar{p}_\ell \bar{p}_k (\bar{D}_\ell^\top \otimes \bar{D}_k) \quad (94)$$

Noting that F_{sync} is identical to F' in (179) of Part II [3], then from Lemma 4 of Part II [3] we obtain

$$\rho(F_{\text{async}}) = \rho(F_{\text{sync}}) + O(\nu^2) \quad (95)$$

Using (92), (93), and (95), we get

$$\begin{aligned} |\rho(\mathcal{F}_{\text{async}}) - \rho(\mathcal{F}_{\text{sync}})| &= |\rho(F_{\text{async}}) - \rho(F_{\text{sync}}) + O(\nu^{1+1/N^2})| \\ &= |O(\nu^2) + O(\nu^{1+1/N^2})| \\ &= O(\nu^{1+1/N^2}) \end{aligned} \quad (96)$$

Using (83) from Part II [3] and (96) completes the proof. \blacksquare

Likewise, assuming $c_{\mu,k,k} = 0$ and $c_{a,\ell k,\ell k} = 0$ for all k and ℓ for the synchronous strategy, it is easy to verify from (77)–(80) of Part II [3] that $p = \bar{p} \otimes \bar{p}$. Then, we obtain the following expression for the steady-state MSD of the synchronous network with the constant parameters $\{\bar{\mu}_k\}$ and $\{\bar{a}_{\ell k}\}$:

$$\text{MSD}_{\text{sync}}^{\text{dist}} = \frac{1}{4} \text{Tr}(H^{-1} R_{\text{sync}}) + O(\nu^{1+\gamma_o}) \quad (97)$$

where $H = O(\nu)$ and $0 < \gamma_o \leq 1/2$ are given by (84) and (70) from Part II [3], respectively, and

$$R_{\text{sync}} \triangleq \sum_{k=1}^N \bar{p}_k^2 \bar{\mu}_k^2 R_k = O(\nu^2) \quad (98)$$

Since $\text{Tr}(H^{-1}R_{\text{sync}}) = O(\nu)$, the first term on the RHS of (97) dominates the other term, $O(\nu^{1+\gamma_o})$. From (86) and (97), we observe that the network MSDs of asynchronous and synchronous networks are both in the order of ν .

Lemma 8 (Degradation in MSD is $O(\nu)$): The network MSD (86) for the asynchronous diffusion strategy is greater than the network MSD (97) for the synchronous diffusion strategy by a difference in the order of ν .

Proof: The difference between R_{async} and R_{sync} is

$$R_{\text{async}} - R_{\text{sync}} = \sum_{k=1}^N [(p_{k,k} - \bar{p}_k^2) \bar{\mu}_k^2 + p_{k,k} c_{\mu,k,k}] R_k \quad (99)$$

where R_{async} is given by (88) of Part II [3]. Since $p_{k,k} - \bar{p}_k^2$ is the k -th entry on the diagonal of $P_p - \bar{p}\bar{p}^T$, from Lemma 3, we know that all entries on the diagonal of $P_p - \bar{p}\bar{p}^T$ are nonnegative, which implies that $p_{k,k} - \bar{p}_k^2 \geq 0$. Moreover, by Perron-Frobenius Theorem [18], all entries of the Perron eigenvector p must be positive, which implies that $p_{k,k} > 0$. We also know that $c_{\mu,k,k}$ must be positive in the asynchronous model. Therefore, we get

$$(p_{k,k} - \bar{p}_k^2) \bar{\mu}_k^2 + p_{k,k} c_{\mu,k,k} > 0 \quad (100)$$

Moreover, by using (186)–(188) from Part II [3], we have

$$(p_{k,k} - \bar{p}_k^2) \bar{\mu}_k^2 + p_{k,k} c_{\mu,k,k} = O(\nu^2) \quad (101)$$

Then, using the fact that the $\{R_k\}$ are positive semi-definite, we conclude from (99)–(101) that

$$\|R_{\text{async}} - R_{\text{sync}}\| = O(\nu^2) > 0 \quad (102)$$

From (84) of Part II [3], we know that $H^{-1} = O(\nu^{-1})$. Therefore, we get

$$\begin{aligned} \text{MSD}_{\text{async}}^{\text{dist}} - \text{MSD}_{\text{sync}}^{\text{dist}} &= \frac{1}{4} \text{Tr}[H^{-1}(R_{\text{async}} - R_{\text{sync}})] + O(\nu^{1+\gamma_o}) \\ &= O(\nu) + O(\nu^{1+\gamma_o}) = O(\nu) \end{aligned} \quad (103)$$

and

$$\text{MSD}_{\text{async}}^{\text{dist}} - \text{MSD}_{\text{sync}}^{\text{dist}} \geq 0 \quad (104)$$

which complete the proof. ■

We observe from the above results that when the step-sizes are sufficiently small, the mean-square convergence rate of the asynchronous network tends to be immune from the uncertainties caused by random topologies, links, agents, and data arrival time. However, there is an $O(\nu)$ degradation in the steady-state MSD level for the asynchronous network – refer to Table I for a summary of the main conclusions.

VI. A CASE STUDY: MSE ESTIMATION

The previous results apply to arbitrary strongly-convex costs $\{J_k(w)\}$ whose Hessian functions are locally Lipschitz continuous at w^o . In this section we specialize the results to the case of MSE estimation over networks, where the costs $\{J_k(w)\}$ become quadratic in $w \in \mathbb{C}^{M \times 1}$.

A. Problem Formulation and Modeling

We now assume that each agent k has access to streaming data $\{\mathbf{d}_k(i), \mathbf{u}_{k,i}\}$ related via the linear regression model:

$$\mathbf{d}_k(i) = \mathbf{u}_{k,i} w^o + \boldsymbol{\xi}_k(i) \quad (105)$$

where $\mathbf{d}_k(i) \in \mathbb{C}$ is the observation, $\mathbf{u}_{k,i} \in \mathbb{C}^{1 \times M}$ is the regressor, $w^o \in \mathbb{C}^{M \times 1}$ is the desired parameter vector, and $\boldsymbol{\xi}_k(i)$ is additive noise.

Assumption 2 (Data model):

- 1) The regressors $\{\mathbf{u}_{k,i}\}$ are temporally white and spatially independent circular symmetric complex random variables with zero mean and covariance matrix $R_{u,k} > 0$.
- 2) The noise signals $\{\boldsymbol{\xi}_k(i)\}$ are temporally white and spatially independent circular symmetric complex random variables with zero mean and variance $\sigma_{\xi,k}^2 > 0$.
- 3) The random variables $\{\mathbf{u}_{k,i}, \boldsymbol{\xi}_\ell(j)\}$ are mutually independent for any k and ℓ , i and j , and they are independent of any other random variable.

The objective for the network is to estimate w^o by minimizing the aggregate mean-square-error cost defined by

$$\underset{w}{\text{minimize}} \sum_{k=1}^N J_k(w) \triangleq \sum_{k=1}^N \mathbb{E} |\mathbf{d}_k(i) - \mathbf{u}_{k,i} w|^2 \quad (106)$$

It can be verified that this problem satisfies Assumptions 1 and 2 introduced in Part I [2].

B. Distributed Diffusion Solutions

The asynchronous diffusion solution (71a)–(71b) will then reduce to the following form:

$$\psi_{k,i} = \mathbf{w}_{k,i-1} + \mu_k(i) \mathbf{u}_{k,i}^* [\mathbf{d}_k(i) - \mathbf{u}_{k,i} \mathbf{w}_{k,i-1}] \quad (107a)$$

$$\mathbf{w}_{k,i} = \sum_{\ell \in \mathcal{N}_{k,i}} a_{\ell k}(i) \psi_{\ell,i} \quad (107b)$$

and the synchronous network (88a)–(88b) will become

$$\psi_{k,i} = \mathbf{w}_{k,i-1} + \bar{\mu}_k \mathbf{u}_{k,i}^* [\mathbf{d}_k(i) - \mathbf{u}_{k,i} \mathbf{w}_{k,i-1}] \quad (108a)$$

$$\mathbf{w}_{k,i} = \sum_{\ell \in \mathcal{N}_{k,i}} \bar{a}_{\ell k} \psi_{\ell,i} \quad (108b)$$

We assume that the network is under the Bernoulli model described in Part I [2]. For illustration purposes only, we assume that the parameters $\{\mu_k\}$ in (55) of Part I [2] are uniform, $\mu_k \equiv \mu$, and that the parameters $\{a_{\ell k}; \ell \in \mathcal{N}_k \setminus \{k\}\}$ in (56) of Part I [2] are given by $a_{\ell k} = |\mathcal{N}_k|^{-1}$.

Substituting (105) into (107a) and comparing with (18) of Part I [2], we find that the approximate gradient, $\widehat{\nabla_{w^*} J_k}(\mathbf{w}_{k,i-1})$, and the corresponding gradient noise, $\mathbf{v}_{k,i}(\mathbf{w}_{k,i-1})$, in this case are given by

$$\begin{aligned} \widehat{\nabla_{w^*} J_k}(\mathbf{w}_{k,i-1}) &= -\mathbf{u}_{k,i}^* [\mathbf{d}_k(i) - \mathbf{u}_{k,i} \mathbf{w}_{k,i-1}] \\ &= -\mathbf{u}_{k,i}^* \mathbf{u}_{k,i} \tilde{\mathbf{w}}_{k,i-1} - \mathbf{u}_{k,i}^* \boldsymbol{\xi}_k(i) \\ &= -R_{u,k} \tilde{\mathbf{w}}_{k,i-1} - \mathbf{v}_{k,i}(\mathbf{w}_{k,i-1}) \end{aligned} \quad (109)$$

where

$$\mathbf{v}_{k,i}(\mathbf{w}_{k,i-1}) = (\mathbf{u}_{k,i}^* \mathbf{u}_{k,i} - R_{u,k}) \tilde{\mathbf{w}}_{k,i-1} + \mathbf{u}_{k,i}^* \boldsymbol{\xi}_k(i) \quad (110)$$

It can be verified that the gradient noise $\mathbf{v}_{k,i}(\mathbf{w}_{k,i-1})$ in (110) satisfies Assumption 1 of Part II [3] and that the covariance matrix of $\mathbf{v}_{k,i}(w^o) = \mathbb{T}(\mathbf{u}_{k,i}^* \boldsymbol{\xi}_k(i))$, where $\mathbb{T}(\cdot)$ is from (4) of Part I [2], is given by

$$R_k = \text{diag}\{\sigma_{\xi,k}^2 R_{u,k}, \sigma_{\xi,k}^2 R_{u,k}^\top\} \triangleq \sigma_{\xi,k}^2 H_k \quad (111)$$

Moreover, the complex Hessian of the cost $J_k(w)$ is given by

$$\nabla_{ww^*}^2 J_k(w) \triangleq H_k = \text{diag}\{R_{u,k}, R_{u,k}^\top\} \quad (112)$$

We further note that for the Bernoulli network under study,

$$\bar{\mu}_k^{(1)} = q_k \mu, \quad \bar{\mu}_k^{(2)} = q_k \mu^2 \quad (113)$$

Therefore, the parameter $\nu = \mu$ in this case. If μ is small enough and satisfies Assumption 1, then from (86), the network MSD of the asynchronous network is given by

$$\text{MSD}_{\text{async}}^{\text{diff}} = \frac{\mu}{2} \text{Tr} \left[\left(\sum_{k=1}^N \bar{p}_k q_k R_{u,k} \right)^{-1} \left(\sum_{k=1}^N p_{k,k} q_k \sigma_{\xi,k}^2 R_{u,k} \right) \right] + O(\mu^{1+\gamma_o}) \quad (114)$$

Likewise, the network MSD of the synchronous network from (97) is given by

$$\text{MSD}_{\text{sync}}^{\text{diff}} = \frac{\mu}{2} \text{Tr} \left[\left(\sum_{k=1}^N \bar{p}_k q_k R_{u,k} \right)^{-1} \left(\sum_{k=1}^N \bar{p}_k^2 q_k^2 \sigma_{\xi,k}^2 R_{u,k} \right) \right] + O(\mu^{1+\gamma_o}) \quad (115)$$

Clearly, since $q_k \leq 1$ and $p_{k,k} \geq \bar{p}_k^2$ for all k , the MSD in (114) is always greater than the MSD in (115) and the difference is in the order of μ .

C. Centralized Solution

The asynchronous batch solution (7) will now reduce to

$$\mathbf{w}_{c,i} = \mathbf{w}_{c,i-1} + \sum_{k=1}^N \pi_k(i) \boldsymbol{\mu}_k(i) \mathbf{u}_{k,i}^* [\mathbf{d}_k(i) - \mathbf{u}_{k,i} \mathbf{w}_{c,i-1}] \quad (116)$$

and the synchronous batch solution (66) will become

$$\mathbf{w}_{c,i} = \mathbf{w}_{c,i-1} + \sum_{k=1}^N \bar{\pi}_k \bar{\mu}_k \mathbf{u}_{k,i}^* [\mathbf{d}_k(i) - \mathbf{u}_{k,i} \mathbf{w}_{c,i-1}] \quad (117)$$

We continue to assume that the random step-size parameters $\{\boldsymbol{\mu}_k(i)\}$ satisfy the same Bernoulli model described in Part I [2] with a uniform profile $\mu_k \equiv \mu$. We use the procedure described in Section IV-B to generate the random fusion coefficients $\{\pi_k(i)\}$. Specifically, we have $\pi_k(i) = \phi_k(i)$, where $\phi_k(i)$ denotes the k -th entry of ϕ_i from (79).

D. Simulation Results

We consider a network consisting of $N = 100$ agents with the connected topology shown in Fig. 1 where each link is assumed to be bidirectional. The length of the unknown parameter w^o is set to $M = 2$. The regressors are assumed to be white, i.e., $R_{u,k} = \sigma_{u,k}^2 I_M$. The values of $\{\sigma_{u,k}^2, \sigma_{v,k}^2\}$ are randomly generated and shown in Fig. 2. The step-size parameter is set to $\mu = 0.002$. We randomly select the values for the probabilities $\{\eta_{\ell k}\}$ in (56) of Part I [2] within the range $(0.4, 0.8)$, and randomly select the values for the probabilities $\{q_k\}$ in (55) of Part I [2] within the set $\{0.3, 0.5, 0.7, 0.9\}$. The asynchronous distributed strategy (107a)–(107b), the synchronous distributed strategy (108a)–(108b), the asynchronous centralized solution (116), and the synchronous centralized solution (117) are all simulated over 100 trials

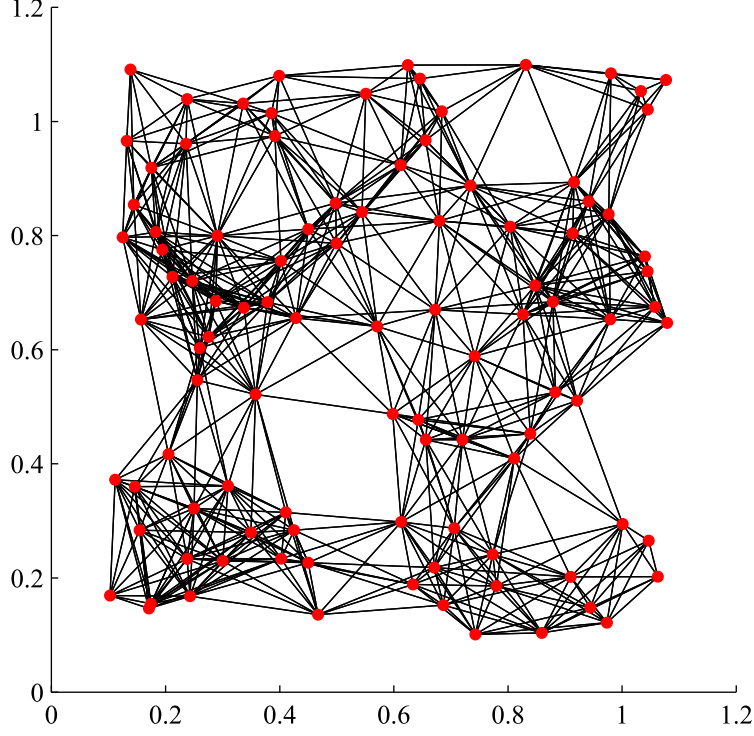


Fig. 1. A topology with 100 nodes.

and 6000 iterations for each trial. The random fusion coefficients $\{\pi_k(i)\}$ are obtained by sampling ϕ_i from (79). The ϕ_i is constructed by consecutively multiplying 100 independent realizations of A_i . The averaged learning curves (MSD) as well as the theoretical MSD results (114) for asynchronous solutions and (115) for synchronous solutions are plotted in Fig. 3. We observe a good match between theory and simulation. We also observe that both synchronous and asynchronous solutions converge at a similar rate but that the former attains a lower MSD level at steady-state as predicted by (114) and (115).

VII. CONCLUSION

In this part, we compared the performance of distributed and centralized solutions under two modes of operation: synchronous and asynchronous implementations. We derived explicit comparisons for the mean and mean-square rates of convergence, as well as for the steady-state mean-square error performance. The main results are captured by Table 1. It is seen that diffusion networks are remarkably resilient to asynchronous or random failures: the convergence continues to occur at the same rate as synchronous or centralized solutions while the MSD level suffers a degradation in the order of $O(\nu)$ relative to synchronous diffusion networks. The results in the article highlight yet another benefit of cooperation: remarkable resilience to random failures and asynchronous events.

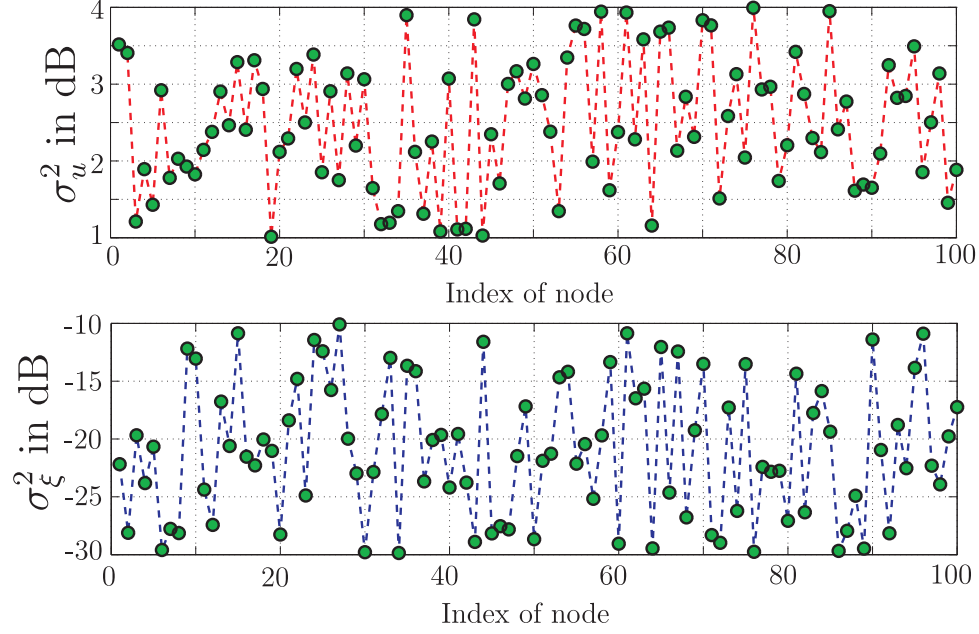


Fig. 2. Values of $\{\sigma_{u,k}^2\}$ and $\{\sigma_{\xi,k}^2\}$.

APPENDIX A

PROOF OF LEMMA 2

From (41), we get

$$\begin{aligned}
 F_c &= \mathbb{E} \left[\left(\sum_{\ell=1}^N \pi_{\ell}(i) \mathbf{D}_{\ell,i} \right)^{\top} \otimes \left(\sum_{k=1}^N \pi_k(i) \mathbf{D}_{k,i} \right) \right] \\
 &\stackrel{(a)}{=} \sum_{k=1}^N \sum_{\ell=1}^N \mathbb{E}[\pi_{\ell}(i) \pi_k(i)] \cdot \mathbb{E}(\mathbf{D}_{\ell,i}^{\top} \otimes \mathbf{D}_{k,i}) \\
 &\stackrel{(b)}{=} \sum_{k=1}^N \sum_{\ell=1}^N (\bar{\pi}_{\ell} \bar{\pi}_k + c_{\pi,\ell,k}) (\bar{\mathbf{D}}_{\ell}^{\top} \otimes \bar{\mathbf{D}}_k + c_{\mu,\ell,k} \mathbf{H}_{\ell}^{\top} \otimes \mathbf{H}_k)
 \end{aligned} \tag{118}$$

where step (a) is by using the independence condition from the asynchronous model; and step (b) is by using (13)–(16). Since $\{\bar{\mathbf{D}}_k, \mathbf{H}_k\}$ are all Hermitian, it is straightforward to verify that F_c is also Hermitian.

Using Jensen's inequality and the convexity of the 2-induced norm, $\|\cdot\|$, we obtain from (51) that

$$\rho(F_c) \leq \mathbb{E} \|\mathbf{B}_i^{\top} \otimes \mathbf{B}_i\| = \mathbb{E} \|\mathbf{B}_i\|^2 \tag{119}$$

where we used the identities $\|A \otimes B\| = \|A\| \cdot \|B\|$ [35, p. 245] and $\|A^{\top}\| = \|A\|$. Using Jensen's inequality again with respect to the convex coefficients $\{\pi_k(i)\}$ and the fact that $\|\cdot\|^2$ is also a convex

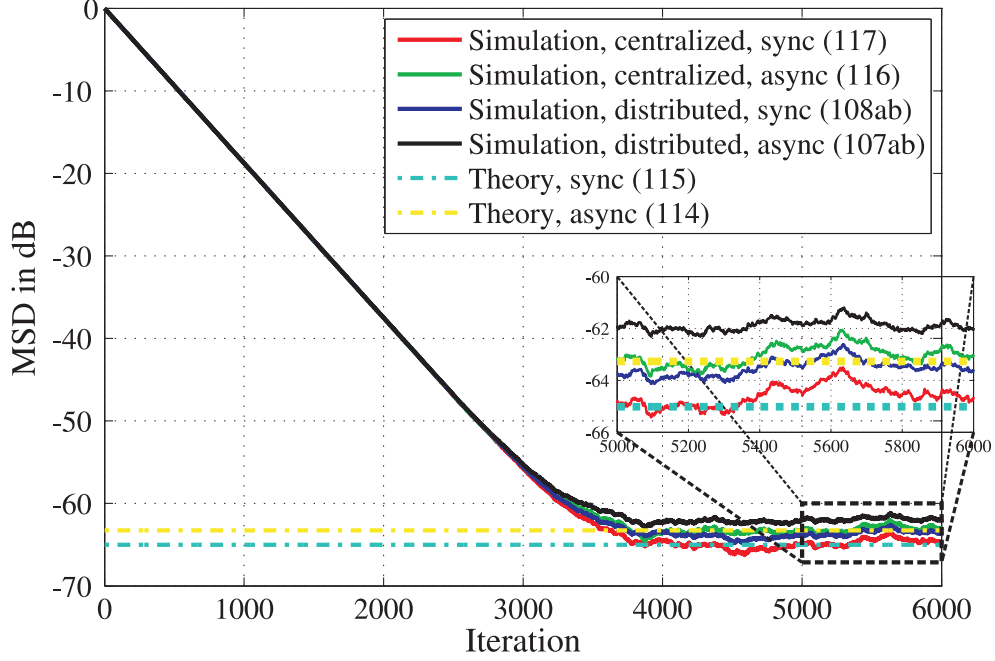


Fig. 3. MSD learning curves for the asynchronous and synchronous modes of operation.

function, we get from (41) that

$$\|B_i\|^2 = \left\| \sum_{k=1}^N \pi_k(i) D_{k,i} \right\|^2 \leq \sum_{k=1}^N \pi_k(i) \|D_{k,i}\|^2 \quad (120)$$

Substituting (120) into (119), we obtain

$$\rho(F_c) \leq \sum_{k=1}^N \bar{\pi}_k \mathbb{E} \|D_{k,i}\|^2 \leq \max_k \mathbb{E} \|D_{k,i}\|^2 \quad (121)$$

From (41) and from condition (8) in Part I [2], we have

$$1 - \mu_k(i) \lambda_{k,\max} \leq \lambda(D_{k,i}) \leq 1 - \mu_k(i) \lambda_{k,\min} \quad (122)$$

for every eigenvalue of $D_{k,i}$ and for every k and $i \geq 0$. Since $D_{k,i}$ is Hermitian, we conclude from (122) that for every k and $i \geq 0$,

$$\begin{aligned} \|D_{k,i}\|^2 &\leq \max\{[1 - \mu_k(i) \lambda_{k,\min}]^2, [1 - \mu_k(i) \lambda_{k,\max}]^2\} \\ &\leq 1 - 2\mu_k(i) \lambda_{k,\min} + \mu_k^2(i) \lambda_{k,\max}^2 \end{aligned} \quad (123)$$

Substituting (123) into (121) yields

$$\begin{aligned} \rho(F_c) &\leq \max_k \mathbb{E} [1 - 2\mu_k(i) \lambda_{k,\min} + \mu_k^2(i) \lambda_{k,\max}^2 | \tilde{\mathbf{w}}_{c,i-1}] \\ &\leq \max_k \{1 - 2\bar{\mu}_k \lambda_{k,\min} + (\bar{\mu}_k^2 + c_{\mu,k,k}) \lambda_{k,\max}^2\} \end{aligned}$$

$$\begin{aligned}
&< \max_k \{\gamma_k^2 + \alpha(\bar{\mu}_k^2 + c_{\mu,k,k})\} \\
&= \beta
\end{aligned} \tag{124}$$

where $\alpha > 0$ and $\{\gamma_k^2, \beta\}$ are from (89) and (90) of Part I [2], respectively. In (144) from Part I [2], we established that $|\beta| < 1$ if condition (33) holds. Therefore, by (124) we conclude that $\rho(F_c) < 1$ when condition (33) holds.

Since $c_{\mu,\ell,k} = O(\nu^2)$ by using (187) and (188) from Part II [3], we get from (118) and (46) that

$$F_c = \sum_{k=1}^N \sum_{\ell=1}^N (\bar{\pi}_\ell \bar{\pi}_k + c_{\pi,\ell,k}) (\bar{D}_\ell^\top \otimes \bar{D}_k) + O(\nu^2) \tag{125}$$

Furthermore, we have

$$\begin{aligned}
\sum_{k=1}^N \sum_{\ell=1}^N c_{\pi,\ell,k} (\bar{D}_\ell^\top \otimes \bar{D}_k) &\stackrel{(a)}{=} \sum_{k=1}^N \sum_{\ell=1}^N c_{\pi,\ell,k} (I_{2M} - \bar{\mu}_\ell H_\ell)^\top \otimes (I_{2M} - \bar{\mu}_k H_k) \\
&= \sum_{k=1}^N \sum_{\ell=1}^N c_{\pi,\ell,k} (I_{4M^2} - \bar{\mu}_\ell H_\ell^\top \otimes I_{2M} - I_{2M} \otimes \bar{\mu}_k H_k + \bar{\mu}_\ell \bar{\mu}_k H_\ell^\top \otimes H_k) \\
&\stackrel{(b)}{=} \sum_{k=1}^N \sum_{\ell=1}^N c_{\pi,\ell,k} \bar{\mu}_\ell \bar{\mu}_k (H_\ell^\top \otimes H_k) \\
&\stackrel{(c)}{=} O(\nu^2)
\end{aligned} \tag{126}$$

where step (a) is by using (47); step (b) is by using (20); and step (c) is by using (187) and (188) from Part II [3]. From (125) and (126), we have

$$F_c = \sum_{\ell=1}^N \sum_{k=1}^N \bar{\pi}_\ell \bar{\pi}_k (\bar{D}_\ell^\top \otimes \bar{D}_k) + O(\nu^2) \tag{127}$$

Now, consider the matrix F'_c defined in (67); it is easy to verify by using (46) that

$$F'_c = \sum_{\ell=1}^N \sum_{k=1}^N \bar{\pi}_\ell \bar{\pi}_k (\bar{D}_\ell^\top \otimes \bar{D}_k) = \bar{B}^\top \otimes \bar{B} \tag{128}$$

Since \bar{B} is Hermitian, so is F'_c . From (127) and (128), we get $\|F_c - F'_c\| = O(\nu^2)$. Since both F_c and F'_c are Hermitian, their difference $F_c - F'_c$ is also Hermitian. Then, using a corollary of the Wielandt-Hoffman Theorem [36], we conclude that

$$|\lambda_m(F_c) - \lambda_m(F'_c)| \leq \|F_c - F'_c\| = O(\nu^2) \tag{129}$$

where $\lambda_m(\cdot)$ denotes the m -th eigenvalue of its Hermitian matrix argument; the eigenvalues are assumed to be ordered from largest to smallest in each case. From (129), we immediately deduce that

$$|\rho(F_c) - \rho(F'_c)| \leq O(\nu^2) \tag{130}$$

From (46)–(47) and (53), we have

$$\bar{B} = I_{2M} - H_c, \quad \lambda(\bar{B}) = 1 - \lambda(H_c) \quad (131)$$

Since H_c is symmetric positive definite, and since the $\{\bar{\pi}_k\}$ are convex coefficients by (20), we get from Jensen's inequality that

$$0 < \lambda(H_c) \leq \|H_c\| \leq \sum_{k=1}^N \bar{\pi}_k \|\bar{\mu}_k H_k\| \leq \max_k \{\bar{\mu}_k \lambda_{k,\max}\} \quad (132)$$

for all eigenvalues of H_c . When condition (33) holds, we have

$$\bar{\mu}_k \leq \bar{\mu}_k(1 + \rho_k^2) < \frac{\lambda_{k,\min}}{\alpha + \lambda_{k,\max}^2} < \frac{1}{\lambda_{k,\max}} \quad (133)$$

for any k . This implies that $\max_k \{\bar{\pi}_k \lambda_{k,\max}\} < 1$ and therefore, $0 < \lambda(H_c) < 1$ for all eigenvalues of H_c . From (186) of Part II [3] and (132), we get

$$0 < \lambda(H_c) = O(\nu) < 1 \quad (134)$$

for any eigenvalue of H_c . Therefore, we get from (131) that

$$\lambda(\bar{B}) = 1 - O(\nu), \quad \rho(\bar{B}) = 1 - \lambda_{\min}(H_c) \quad (135)$$

Then, from (128) and (135), we have

$$\rho(F'_c) = [1 - \lambda_{\min}(H_c)]^2 \quad (136)$$

It then follows from (130) and (136) that

$$\rho(F_c) = [1 - \lambda_{\min}(H_c)]^2 + O(\nu^2) \quad (137)$$

where $\lambda_{\min}(H_c) = O(\nu)$. Under Assumption 1, we have

$$[1 - \lambda_{\min}(H_c)]^2 = 1 - 2\lambda_{\min}(H_c) + O(\nu^2) = 1 - O(\nu) \quad (138)$$

which therefore dominates the $O(\nu^2)$ in (137).

From (131) and (128), we get

$$F'_c = I_{4M^2} - H_c^\top \otimes I_{2M} - I_{2M} \otimes H_c + H_c^\top \otimes H_c \quad (139)$$

Then, using (127), (128), and (139), we have

$$I_{4M^2} - F_c = \underbrace{H_c^\top \otimes I_{2M} + I_{2M} \otimes H_c}_{= O(\nu)} + O(\nu^2) \quad (140)$$

where we used the fact that $H_c^\top \otimes H_c = O(\nu^2)$ since $H_c = O(\nu)$ by (53). Using the fact that H_c is positive definite and is of the order of ν , we eventually get

$$\|(I_{4M^2} - F_c)^{-1}\| = O(\nu^{-1}) \quad (141)$$

APPENDIX B
PROOF OF THEOREM 5

We start with the $\lim_{i \rightarrow \infty} y_{c,i}$ in (59). From (52), we have

$$\lim_{i \rightarrow \infty} y_{c,i} = \lim_{i \rightarrow \infty} \text{vec}(\mathbb{E} \underline{\mathbf{s}}_i \underline{\mathbf{s}}_i^*) \quad (142)$$

Using the gradient noise model from Section 1 of Part II [3], it can be verified that $\underline{\mathbf{s}}_i$ is zero mean and that its conditional covariance matrix is given by

$$\begin{aligned} \mathbb{E}[\underline{\mathbf{s}}_i \underline{\mathbf{s}}_i^* | \mathbb{F}_{i-1}] &\stackrel{(a)}{=} \sum_{\ell=1}^N \sum_{k=1}^N \mathbb{E}[\boldsymbol{\pi}_\ell(i) \boldsymbol{\pi}_k(i)] \cdot \mathbb{E}[\boldsymbol{\mu}_\ell(i) \boldsymbol{\mu}_k(i)] \mathbb{E}[\mathbf{v}_{\ell,i}(\mathbf{w}_{c,i-1}) \mathbf{v}_{k,i}^*(\mathbf{w}_{c,i-1}) | \mathbb{F}_{i-1}] \\ &\stackrel{(b)}{=} \sum_{k=1}^N (\bar{\pi}_k^2 + c_{\pi,k,k})(\bar{\mu}_k^2 + c_{\mu,k,k}) R_{k,i}(\mathbf{w}_{c,i-1}) \end{aligned} \quad (143)$$

where step (a) is by using the independence condition from the asynchronous model in Part I [2]; and step (b) is from (19) in Part I [2], (7) in Part II [3], and (13)–(16). Therefore,

$$\mathbb{E} \underline{\mathbf{s}}_i \underline{\mathbf{s}}_i^* = \sum_{k=1}^N (\bar{\pi}_k^2 + c_{\pi,k,k})(\bar{\mu}_k^2 + c_{\mu,k,k}) \mathbb{E} R_{k,i}(\mathbf{w}_{c,i-1}) \quad (144)$$

Note that

$$\begin{aligned} \|R_{k,i}(w^o) - \mathbb{E} R_{k,i}(\mathbf{w}_{c,i-1})\| &\stackrel{(a)}{\leq} \|\mathcal{R}_i(\mathbf{1}_N \otimes w^o) - \mathbb{E} \mathcal{R}_i(\mathbf{1}_N \otimes \mathbf{w}_{c,i-1})\| \\ &\stackrel{(b)}{\leq} \kappa_v \cdot [\mathbb{E} \|\mathbf{1}_N \otimes \tilde{\mathbf{w}}_{c,i-1}\|^4]^{\gamma_v/4} \\ &\stackrel{(c)}{=} \kappa_v N^{\gamma_v/2} \cdot [\mathbb{E} \|\tilde{\mathbf{w}}_{c,i-1}\|^4]^{\gamma_v/4} \end{aligned} \quad (145)$$

where step (a) is due to (7) from Part II [3]; step (b) is by using (60) also from Part II [3]; and step (c) is by the fact that $\|\mathbf{1}_N \otimes x\|^4 = [N \cdot \|x\|^2]^2 = N^2 \cdot \|x\|^4$ for any x . Under Assumption 1, we can get from Theorem 2 that

$$\limsup_{i \rightarrow \infty} \|R_{k,i}(w^o) - \mathbb{E} R_{k,i}(\mathbf{w}_{c,i-1})\| \leq \kappa_v N^{\gamma_v/2} \cdot [b_4^2 \nu^2]^{\gamma_v/4} = O(\nu^{\gamma_v/2}) \quad (146)$$

which means that, asymptotically, we can replace $\mathbb{E} R_{k,i}(\mathbf{w}_{c,i-1})$ by R_k from (10) of Part II [3] within an error in the order of $\nu^{\gamma_v/2}$. Therefore, it follows from (142) that

$$\begin{aligned} \lim_{i \rightarrow \infty} y_{c,i} &= \text{vec} \left(\lim_{i \rightarrow \infty} \mathbb{E} \underline{\mathbf{s}}_i \underline{\mathbf{s}}_i^* \right) \\ &\stackrel{(a)}{=} \text{vec} \left(\sum_{k=1}^N (\bar{\pi}_k^2 + c_{\pi,k,k})(\bar{\mu}_k^2 + c_{\mu,k,k}) \lim_{i \rightarrow \infty} \mathbb{E} R_{k,i}(\mathbf{w}_{c,i-1}) \right) \\ &\stackrel{(b)}{=} \text{vec} \left(\sum_{k=1}^N (\bar{\pi}_k^2 + c_{\pi,k,k})(\bar{\mu}_k^2 + c_{\mu,k,k}) [R_k + O(\nu^{\gamma_v/2})] \right) \end{aligned}$$

$$\stackrel{(c)}{=} \text{vec}(R_c) + O(\nu^{2+\gamma_v/2}) \quad (147)$$

where step (a) is by using (144); step (b) is by using (146); and step (c) is by using (63) and the fact from (198) of Part I [2] that $\bar{\mu}_k^2 + c_{\mu,k,k} = \bar{\mu}_k^{(2)} = O(\nu^2)$. Substituting (147) into (59) yields

$$z_{c,\infty} = (I_{4M^2} - F_c)^{-1} \cdot \text{vec}(R_c) + O(\nu^{1+\gamma_v/2}) \quad (148)$$

where we used Lemma 2. Substituting (59) and $\Sigma = I_{2M}$ into (60), and using (56) as well as the fact that F_c and R_c are Hermitian, we obtain

$$\lim_{i \rightarrow \infty} \mathbb{E} \|\tilde{\mathbf{w}}'_{c,i}\|^2 = \frac{1}{2} [\text{vec}(R_c)]^* (I_{4M^2} - F_c)^{-1} \text{vec}(I_{2M}) + O(\nu^{1+\gamma_v/2}) \quad (149)$$

Substituting (149) into (62) yields (64).

We establish (65) next. From (140), we know that

$$I_{4M^2} - F_c = S_c + O(\nu^2) \quad (150)$$

where

$$S_c \triangleq H_c^\top \otimes I_{2M} + I_{2M} \otimes H_c = O(\nu) \quad (151)$$

Since H_c is symmetric and positive definite by (134), it is easy to verify that S_c is also symmetric and positive definite. Therefore, S_c is invertible. Using the matrix inversion lemma [37], we get from (150) that

$$(I_{4M^2} - F_c)^{-1} = S_c^{-1} + O(1) \quad (152)$$

where we used the fact that $\|S_c^{-1}\| = O(\nu^{-1})$. Substituting (152) into (64) yields:

$$\begin{aligned} \text{MSD}^{\text{cent}} &= \frac{1}{2} [\text{vec}(R_c)]^* [S_c^{-1} + O(1)] \text{vec}(I_{2M}) + O(\nu^{1+\gamma_o}) \\ &= \frac{1}{2} [\text{vec}(R_c)]^* S_c^{-1} \text{vec}(I_{2M}) + O(\nu^2) + O(\nu^{1+\gamma_o}) \\ &= \frac{1}{2} [\text{vec}(R_c)]^* S_c^{-1} \text{vec}(I_{2M}) + O(\nu^{1+\gamma_o}) \end{aligned} \quad (153)$$

where we used the fact from (63) that $\|R_c\| = O(\nu^2)$ and $\gamma_o < 1/2$ from (70) of Part II [3]. Since the first term on the RHS of (153) is of the order of ν , it is the dominant term under Assumption 1. To further simplify (153), we introduce the Lyapunov equation with respect to the unknown square matrix X :

$$XH_c + H_cX = I_{2M} \quad (154)$$

where H_c is given by (53). Vectorizing both sides and using (151), the Lyapunov equation is equivalent to the linear system of equations:

$$S_c \text{vec}(X) = \text{vec}(I_{2M}) \quad (155)$$

Since S_c is invertible, the linear equation (155) has a unique solution, which is given by $X = \frac{1}{2}H_c^{-1}$.

From the Lyapunov equation (154) we get

$$\begin{aligned} [\text{vec}(R_c)]^* S_c^{-1} \text{vec}(I_{2M}) &= \frac{1}{2} [\text{vec}(R_c)]^* \text{vec}(H_c^{-1}) \\ &= \frac{1}{2} \text{Tr}(H_c^{-1} R_c) \end{aligned} \quad (156)$$

where we used the fact that R_c is Hermitian. Result (65) then follows from (153) and (156). The term $\text{Tr}(H_c^{-1} R_c) = O(\nu)$ in (65) is the dominant term under Assumption 1.

APPENDIX C

PROOF OF LEMMA 3

From Lemma 3 of Part II [3], we know that P_p is symmetric and, therefore, the matrix difference $C_p \triangleq P_p - \bar{p}\bar{p}^\top$ is also symmetric. We also know from Lemma 3 of Part II [3] that $C_p \mathbb{1}_N = 0$. To establish that C_p is positive semi-definite, we consider the following quadratic expression:

$$x^\top C_p x = x^\top (P_p - \bar{p}\bar{p}^\top) x = x^\top P_p x - (x^\top \bar{p})^2 \quad (157)$$

for any vector $x \in \mathbb{R}^N$. Note that

$$x^\top P_p x = \text{vec}(x^\top P_p x) = \frac{1}{N^2} (x^\top \otimes x^\top) p \cdot \mathbb{1}_{N^2}^\top \mathbb{1}_{N^2} \quad (158)$$

by using the relation $p = \text{vec}(P_p)$ from (80) of Part II [3] and the fact that $\mathbb{1}_{N^2}^\top \mathbb{1}_{N^2} = N^2$. Since

$$\bar{A} \otimes \bar{A} + C_A = \mathbb{E}(\mathbf{A}_j \otimes \mathbf{A}_j) \quad (159)$$

we can introduce a series of fictitious random combination matrices $\{\mathbf{A}'_j; j \geq 1\}$ such that they are mutually-independent and satisfy

$$\mathbb{E}(\mathbf{A}'_j \otimes \mathbf{A}'_j) = \bar{A} \otimes \bar{A} + C_A \quad (160)$$

for any $j \geq 1$. Let $\Phi_i \triangleq \prod_{j=1}^i \mathbf{A}'_j$ for any $i \geq 1$. Then,

$$\lim_{i \rightarrow \infty} \mathbb{E}(\Phi_i \otimes \Phi_i) \stackrel{(a)}{=} \lim_{i \rightarrow \infty} \prod_{j=1}^i \mathbb{E}(\mathbf{A}'_j \otimes \mathbf{A}'_j) \stackrel{(b)}{=} p \cdot \mathbb{1}_{N^2}^\top \quad (161)$$

where step (a) is by using the fact that the $\{\mathbf{A}'_j\}$ are mutually-independent, and step (b) is by using (159) and the Perron-Frobenius Theorem [18]. Substituting (161) into (158) and using $\mathbf{1}_{N^2} = \mathbf{1}_N \otimes \mathbf{1}_N$, we get

$$x^\top P_p x = \frac{1}{N^2} \lim_{i \rightarrow \infty} \mathbb{E}[(x^\top \Phi_i \mathbf{1}_N)^2] \quad (162)$$

Moreover, since $\bar{A} = \mathbb{E}(\mathbf{A}_j | \mathbf{w}_{j-1})$, we have

$$\lim_{i \rightarrow \infty} \mathbb{E}(\Phi_i) = \lim_{i \rightarrow \infty} \prod_{j=1}^i \mathbb{E}(\mathbf{A}'_j) = \lim_{i \rightarrow \infty} (\bar{A})^i = \bar{p} \cdot \mathbf{1}_N^\top \quad (163)$$

Then, using (163) and the fact that $\mathbf{1}_N^\top \mathbf{1}_N = N$, we have

$$x^\top \bar{p} = \frac{1}{N} x^\top \bar{p} \cdot \mathbf{1}_N^\top \mathbf{1}_N = \frac{1}{N} \lim_{i \rightarrow \infty} \mathbb{E}(x^\top \Phi_i \mathbf{1}_N) \quad (164)$$

Substituting (162) and (164) into (157) yields

$$x^\top C_p x = \frac{1}{N^2} \lim_{i \rightarrow \infty} \left\{ \mathbb{E}[(x^\top \Phi_i \mathbf{1}_N)^2] - [\mathbb{E}(x^\top \Phi_i \mathbf{1}_N)]^2 \right\} \geq 0 \quad (165)$$

which confirms that C_p is positive semi-definite.

APPENDIX D

PROOF OF LEMMA 4

We prove Lemma 3 by using a procedure similar to the one given in Appendix I of Part II [3]. Introduce the Jordan decomposition [37]:

$$\bar{A} = \bar{P} \bar{J} \bar{Q}^\top = \begin{bmatrix} \bar{p} & \bar{P}' \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \bar{J}' \end{bmatrix} \begin{bmatrix} \mathbf{1}_N & \bar{Q}' \end{bmatrix}^\top \quad (166)$$

where \bar{J}' is a sub-matrix of \bar{J} containing its stable eigenvalues, \bar{P}' and \bar{Q}' are sub-matrices of \bar{P} and \bar{Q} , and $\bar{P}^{-1} = \bar{Q}^\top$. Then, the Jordan decomposition of $\bar{\mathcal{A}} = \bar{A} \otimes I_{2M}$ from (30) of Part II [3] is given by

$$\bar{\mathcal{A}} = \bar{\mathcal{P}} \bar{\mathcal{J}} \bar{\mathcal{Q}}^\top = \begin{bmatrix} \bar{p}' & \bar{\mathcal{P}}' \end{bmatrix} \begin{bmatrix} I_{2M} & 0 \\ 0 & \bar{\mathcal{J}}' \end{bmatrix} \begin{bmatrix} \bar{q}' & \bar{\mathcal{Q}}' \end{bmatrix}^\top \quad (167)$$

where

$$\bar{\mathcal{P}} = \bar{P} \otimes I_{2M}, \quad \bar{\mathcal{P}}' \triangleq \bar{P}' \otimes I_{2M} \quad (168)$$

$$\bar{\mathcal{J}} = \bar{J} \otimes I_{2M}, \quad \bar{\mathcal{J}}' \triangleq \bar{J}' \otimes I_{2M} \quad (169)$$

$$\bar{\mathcal{Q}} = \bar{Q} \otimes I_{2M}, \quad \bar{\mathcal{Q}}' \triangleq \bar{Q}' \otimes I_{2M} \quad (170)$$

$$\bar{p}' = \bar{p} \otimes I_{2M}, \quad \bar{q}' \triangleq \mathbb{1}_N \otimes I_{2M} \quad (171)$$

Let

$$\bar{\mathcal{X}} \triangleq I_{2MN} - \bar{\mathcal{D}} = \bar{\mathcal{M}}\mathcal{H} = O(\nu) \quad (172)$$

where $\{\bar{\mathcal{D}}, \bar{\mathcal{M}}, \mathcal{H}\}$ are from (33), (31), and (15) of Part II [3], respectively. Then, by (34) from Part II [3] and using the fact that $\bar{\mathcal{A}}$ is real and $\bar{\mathcal{D}}$ is Hermitian, we get

$$\bar{\mathcal{Q}}^T \bar{\mathcal{B}}^* \bar{\mathcal{P}} = \bar{\mathcal{Q}}^T \bar{\mathcal{D}} \bar{\mathcal{A}} \bar{\mathcal{P}} = \begin{bmatrix} I_{2M} - \bar{q}'^T \bar{\mathcal{X}} \bar{p}' & -\bar{q}'^T \bar{\mathcal{X}} \bar{\mathcal{P}}' \bar{\mathcal{J}}' \\ -\bar{\mathcal{Q}}'^T \bar{\mathcal{X}} \bar{p}' & \bar{\mathcal{J}}' - \bar{\mathcal{Q}}'^T \bar{\mathcal{X}} \bar{\mathcal{P}}' \bar{\mathcal{J}}' \end{bmatrix} \quad (173)$$

Using (171) and (172) above and (15) and (31) from Part II [3], we obtain

$$\bar{q}'^T \bar{\mathcal{X}} \bar{p}' = \bar{q}'^T \bar{\mathcal{M}} \mathcal{H} \bar{p}' = \sum_{k=1}^N \bar{p}_k \bar{\mu}_k H_k = H = O(\nu) \quad (174)$$

where H is given by (84) of Part II [3]. By (172), we get

$$\|\bar{q}'^T \bar{\mathcal{X}} \bar{\mathcal{P}}' \bar{\mathcal{J}}'\| = O(\nu), \quad \|\bar{\mathcal{Q}}'^T \bar{\mathcal{X}} \bar{p}'\| = O(\nu), \quad \|\bar{\mathcal{Q}}'^T \bar{\mathcal{X}} \bar{\mathcal{P}}' \bar{\mathcal{J}}'\| = O(\nu) \quad (175)$$

Therefore, we get from (173)–(175) that

$$\bar{\mathcal{Q}}^T \bar{\mathcal{B}}^* \bar{\mathcal{P}} = \begin{bmatrix} \bar{B}_d & O(\nu) \\ O(\nu) & \bar{\mathcal{J}}' + O(\nu) \end{bmatrix} \quad (176)$$

where

$$\bar{B}_d \triangleq I_{2M} - H \quad (177)$$

is Hermitian. From (183) of Part II [3], we immediately get

$$\lambda(\bar{B}_d) = \lambda(I_{2M} - H) = 1 - O(\nu) > 0 \quad (178)$$

$$\rho(\bar{B}_d) = 1 - \lambda_{\min}(H) = 1 - O(\nu) \quad (179)$$

for sufficiently small ν under Assumption 1. Conjugating both sides of (176) and using the fact that \bar{B}_d is Hermitian, we get

$$\bar{B}_s \triangleq (\bar{\mathcal{Q}}^T \bar{\mathcal{B}}^* \bar{\mathcal{P}})^* = \bar{\mathcal{P}}^* \bar{\mathcal{B}} (\bar{\mathcal{Q}}^*)^T = \begin{bmatrix} \bar{B}_d & O(\nu) \\ O(\nu) & \bar{\mathcal{J}}'^* + O(\nu) \end{bmatrix} \quad (180)$$

Since \bar{B}_s is similar to \bar{B} , they have the same eigenvalues [37]. Since \bar{B}_d is Hermitian, let us introduce its eigenvalue decomposition as

$$\bar{B}_d = \bar{U} \bar{\Lambda} \bar{U}^* \quad (181)$$

where \bar{U} is a $2M \times 2M$ unitary matrix and $\bar{\Lambda}$ is a $2M \times 2M$ diagonal matrix. The $(N-1) \times (N-1)$ matrix \bar{J}' , which contains the stable eigenvalues of \bar{A} in (166), can be generally expressed as

$$\bar{J}' = \begin{bmatrix} \bar{\lambda}_{a,2} & & \bar{T}' \\ & \ddots & \\ 0 & & \bar{\lambda}_{a,N} \end{bmatrix} \quad (182)$$

where $\{\bar{\lambda}_{a,n}\}$ are the eigenvalues of \bar{A} with $\bar{\lambda}_{a,1} = 1$ and $|\bar{\lambda}_{a,n}| < 1$ for all $n = 2, 3, \dots, N$. In (182), the elements in the strictly upper triangular region \bar{T}' are either 1 or 0, which depend on the Jordan blocks in \bar{J}' . Using (182) and (169), we can express the $(2, 2)$ block in (180) as

$$\bar{J}'^* + O(\nu) = \begin{bmatrix} \bar{\lambda}_{a,2}^* I_{2M} + O(\nu) & O(\nu) \\ & \ddots \\ \bar{T}'^* + O(\nu) & \bar{\lambda}_{a,N}^* I_{2M} + O(\nu) \end{bmatrix} \quad (183)$$

where the elements in the strictly lower triangular region \bar{T}'^* are either 1 or 0, which depend on the elements of \bar{T}' in (182). We now apply a similarity transformation to \bar{B}_s by multiplying

$$\bar{D} \triangleq \text{diag}\{\nu^\epsilon \bar{U}, \nu^{2\epsilon} I_{2M}, \nu^{3\epsilon} I_{2M}, \dots, \nu^{N\epsilon} I_{2M}\} \quad (184)$$

and its inverse \bar{D}^{-1} on either side of (180), where $\epsilon = 1/N$. Using (180) and (183), we end up with

$$\bar{D} \bar{B}_s \bar{D}^{-1} = \left[\begin{array}{c|ccc} \bar{\Lambda} & & & O(\nu^\epsilon) \\ \hline & \bar{\lambda}_{a,2}^* I_{2M} + O(\nu) & & O(\nu^\epsilon) \\ O(\nu^{1+\epsilon}) & & \ddots & \\ & O(\nu^\epsilon) & & \bar{\lambda}_{a,N}^* I_{2M} + O(\nu) \end{array} \right] \quad (185)$$

From (185), we know that all off-diagonal entries of $\bar{D} \bar{B}_s \bar{D}^{-1}$ are *at least* of the order of ν^ϵ . Therefore, using Gershgorin Theorem [36, p. 320] under Assumption 1, and since \bar{B} and \bar{B}_s have the same eigenvalues due to similarity, we get

$$|\lambda(\bar{B}) - \lambda(\bar{B}_d)| \leq O(\nu^{1+\epsilon}) \quad \text{or} \quad |\lambda(\bar{B}) - \bar{\lambda}_{a,k}^*| \leq O(\nu^\epsilon) \quad (186)$$

where $\lambda(\bar{B})$ denotes the eigenvalue of \bar{B} and $k = 2, 3, \dots, N$. Result (186) implies that the eigenvalues of \bar{B} are either located in the Gershgorin circles that are centered at the eigenvalues of \bar{B}_d with radii $O(\nu^{1+\epsilon})$ or in the Gershgorin circles that are centered at $\{\bar{\lambda}_{a,k}^*; k = 2, 3, \dots, N\}$ with radii $O(\nu^\epsilon)$. From (179), we have

$$\rho(\bar{B}_d) = 1 - O(\nu) < 1 \quad (187)$$

By Assumption 3 from Part II [3] and Perron-Frobenius Theorem [18], we have

$$\rho(\bar{J}^*) \triangleq \max_{k=2,3,\dots,N} |\bar{\lambda}_{a,k}^*| = \rho(\bar{J}') < 1 \quad (188)$$

By Assumption 1, if the parameter ν is small enough such that

$$\rho(\bar{J}') + O(\nu^\epsilon) < 1 - O(\nu) = \rho(\bar{B}_d) \quad (189)$$

holds, then the Gershgorin circles centered at the eigenvalues of \bar{B}_d are isolated from those centered at $\{\bar{\lambda}_{a,k}^*; k = 2, 3, \dots, N\}$. According to Gershgorin Theorem [38, p. 181], there are precisely $2M$ eigenvalues of \bar{B} satisfying

$$|\lambda(\bar{B}) - \lambda(\bar{B}_d)| \leq O(\nu^{1+\epsilon}) \quad (190)$$

while all the other eigenvalues satisfy

$$|\lambda(\bar{B}) - \bar{\lambda}_{a,k}^*| \leq O(\nu^\epsilon), \quad k = 2, 3, \dots, N \quad (191)$$

By (189), the eigenvalues $\lambda(\bar{B})$ satisfying (190) are greater than those satisfying (191) in magnitude. Furthermore, when ν is sufficiently small, the Gershgorin circles centered at $\lambda_{\max}(\bar{B}_d)$ with radius $O(\nu^{1+\epsilon})$ will become disjoint from the other circles. Then, by using Gershgorin Theorem again, we conclude from (190) that

$$|\rho(\bar{B}) - \rho(\bar{B}_d)| \leq O(\nu^{1+\epsilon}) \quad (192)$$

It is worth noting that from (187) and (192) we get

$$\rho(\bar{B}) \leq 1 - O(\nu) + O(\nu^{1+\epsilon}) < 1 \quad (193)$$

for $\nu \ll 1$ because $\epsilon = 1/N > 1$. Eventually, using (46), (73), and (177), it is straightforward to verify that

$$\bar{B} = \bar{B}_d \quad (194)$$

Using (179), (192), and (194) completes the proof.

REFERENCES

- [1] X. Zhao and A. H. Sayed, "Asynchronous diffusion adaptation over networks," in *Proc. European Signal Process. Conf. (EUSIPCO)*, Bucharest, Romania, Aug. 2012, pp. 27–31.
- [2] X. Zhao and A. H. Sayed, "Asynchronous adaptation and learning over networks — Part I: Modeling and stability analysis," *IEEE Trans. Signal Process.*, vol. xx, no. xx, pp. xxxx–xxxx, xxx 2015.
- [3] X. Zhao and A. H. Sayed, "Asynchronous adaptation and learning over networks — Part II: Performance analysis," *IEEE Trans. Signal Process.*, vol. xx, no. xx, pp. xxxx–xxxx, xxx 2015.

- [4] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Autom. Control*, vol. 31, no. 9, pp. 803–812, Sept. 1986.
- [5] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2508–2530, June 2006.
- [6] S. Kar and J. M. F. Moura, "Distributed consensus algorithms in sensor networks: Link failures and channel noise," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 355–369, Jan. 2009.
- [7] K. Srivastava and A. Nedic, "Distributed asynchronous constrained stochastic optimization," *IEEE J. Sel. Top. Signal Process.*, vol. 5, no. 4, pp. 772–790, Aug. 2011.
- [8] S. Kar and J. M. F. Moura, "Sensor networks with random links: Topology design for distributed consensus," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3315–3326, July 2008.
- [9] T. C. Aysal, A. D. Sarwate, and A. G. Dimakis, "Reaching consensus in wireless networks with probabilistic broadcast," in *Proc. Allerton Conf. Commun., Control, Comput.*, Allerton House, IL, Sept. and Oct. 2009, pp. 732–739.
- [10] T. C. Aysal, M. E. Yildiz, and A. Scaglione, "Broadcast gossip algorithms for consensus," *IEEE Trans. Signal Process.*, vol. 57, pp. 2748–2761, 2009.
- [11] S. Kar and J. M. F. Moura, "Distributed consensus algorithms in sensor networks: Quantized data and random link failures," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1383–1400, Mar. 2010.
- [12] D. Jakovetic, J. Xavier, and J. M. F. Moura, "Weight optimization for consensus algorithms with correlated switching topology," *IEEE Trans. Signal Process.*, vol. 58, no. 7, pp. 3788–3801, July 2010.
- [13] S. Kar and J. M. F. Moura, "Convergence rate analysis of distributed gossip (linear parameter) estimation: Fundamental limits and tradeoffs," *IEEE J. Sel. Top. Signal Process.*, vol. 5, no. 4, pp. 674–690, Aug. 2011.
- [14] A. H. Sayed, "Diffusion adaptation over networks," in *Academic Press Library in Signal Processing*, R. Chellapa and S. Theodoridis, Eds., vol. 3, pp. 323–454. Academic Press, Elsevier, 2014.
- [15] A. H. Sayed, S.-Y. Tu, J. Chen, X. Zhao, and Z. Towfic, "Diffusion strategies for adaptation and learning over networks," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 155–171, May 2013.
- [16] A. H. Sayed, "Adaptive networks," *Proc. IEEE*, vol. 102, no. 4, pp. 460–497, Apr. 2014.
- [17] X. Zhao and A. H. Sayed, "Attaining optimal batch performance via distributed processing over networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Vancouver, Canada, May 2013, pp. 1–5.
- [18] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, SIAM, PA, 1994.
- [19] S. U. Pillai, T. Suel, and S. Cha, "The Perron-Frobenius theorem: Some of its applications," *IEEE Signal Process. Mag.*, vol. 22, no. 2, pp. 62–75, Mar. 2005.
- [20] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge Univ. Press, Cambridge, UK, 2004.
- [21] S. Kotz, N. Balakrishnan, and N. L. Johnson, *Continuous Multivariate Distributions Vol. 1: Models and Applications*, Wiley, New York, 2nd edition, 2000.
- [22] R. J. Connor and J. E. Mosimann, "Concepts of independence for proportions with a Generalization of the Dirichlet distribution," *J. Am. Stat. Assoc.*, vol. 64, no. 325, pp. 194–206, Mar. 1969.
- [23] J. Aitchison, "A general class of distributions on the simplex," *J. R. Statist. Soc. B*, vol. 47, no. 1, pp. 136–146, 1985.
- [24] O. E. Barndorff-Nielsen and B. Jorgensen, "Some parametric models on the simplex," *J. Multivariate Anal.*, vol. 39, no. 1, pp. 106–116, Oct. 1991.
- [25] T.-T. Wong, "Generalized Dirichlet distribution in Bayesian analysis," *Appl. Math. Comput.*, vol. 97, no. 2-3, pp. 165–181, Dec. 1998.

- [26] R. K. S. Hankin, “A generalization of the Dirichlet distribution,” *J. Stat. Soft.*, vol. 33, no. 11, pp. 1–18, Feb. 2010.
- [27] W.-Y. Chang, R. D. Gupta, and D. St. P. Richards, “Structural properties of the generalized Dirichlet distributions,” *Contemp. Math.*, vol. 516, pp. 109–124, 2010.
- [28] T.-T. Wong, “Parameter estimation for generalized Dirichlet distributions from the sample estimates of the first and the second moments of random variables,” *Comput. Stat. Data Anal.*, vol. 54, no. 7, pp. 1756–1765, July 2010.
- [29] S. Favaro, G. Hadjicharalambous, and I. Prunster, “On a class of distributions on the simplex,” *J. Stat. Plan Infer.*, vol. 141, no. 9, pp. 2987–3004, Sept. 2011.
- [30] J. Aitchison and S. M. Shen, “Logistic-Normal distributions: Some properties and uses,” *Biometrika*, vol. 67, no. 2, pp. 261–272, Aug. 1980.
- [31] P. J. Lenk, “The Logistic Normal distribution for Bayesian, nonparametric, predictive densities,” *J. Am. Stat. Assoc.*, vol. 83, no. 402, pp. 509–516, June 1988.
- [32] V. Seshadri, “General exponential models on the unit simplex and related multivariate inverse Gaussian distributions,” *Stat. Probabil. Lett.*, vol. 14, no. 5, pp. 385–391, July 1992.
- [33] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, “An introduction to MCMC for machine learning,” *Machine Learning*, vol. 50, no. 1-2, pp. 5–43, Jan. 2003.
- [34] A. Gelman, “Method of moments using Monte Carlo simulation,” *J. Comput. Graph Stat.*, vol. 4, no. 1, pp. 36–54, Feb. 1995.
- [35] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge Univ. Press, Cambridge, UK, 1991.
- [36] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 3rd edition, 1996.
- [37] A. J. Laub, *Matrix Analysis for Scientists and Engineers*, SIAM, PA, 2005.
- [38] G. W. Stewart and J. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, MA, 1990.